Sequence Analysis

Brendan Halpin, University of Limerick

Revisions for 14 June

Contents

1 INTRODUCTION

Sequence analysis in sociology refers to a group of approaches to linear (predominantly longitudinal) data which focus on sequences (such as work-life histories, or conversations) as wholes. Sequence analysis is often exploratory and descriptive in intention, typically oriented to generating data-driven typologies, and can be contrasted with conventional approaches to longitudinal data such as hazard-rate modelling (event history analysis), models of transition patterns, or latent growth curve models, which focus on modelling the processes generating the sequences. The take-up of sequence analysis in sociology has been informed by a perception that complex sequences – such as life-course trajectories, or the rhetorical structure of a story – are likely to have structures that are not easily captured by models that focus on their evolution through time, but may be apparent when they are considered as wholes. Sequences are linear objects, that is, they have a uni-dimensional ordered structure. In sociological research the dimension is almost always time, so sequences are longitudinal. Two main types of sequences exist: those representing longitudinal process such as life-histories, coded as states in successive time-periods, and those representing structures which unfold through time, such as conversations (coded into types of utterances) or dances (coded into sequences of steps). Sequences are very detailed: for instance, a sequence 10 units long in a 4-element state space has more than a million possible forms. Hence, sequence data can be difficult to classify either a priori or by inspection. Sequence analysis often has as goal a data-driven classification, typically achieved by defining a metric of similarity between sequences, and using cluster analysis to group sequences on the basis of pairwise similarity. Sequence analysis draws on computer-science techniques for pattern-finding in strings of tokens (e.g., text) or other longitudinal data (e.g., recorded speech), some of which have proven very powerful, particularly in molecular biology. A recurrent theme in the sequence analysis literature is whether algorithms developed for, or appropriate to, non-sociological fields such as molecular genetics, can map onto sociological data in a meaningful way.

1.1 Andrew Abbott

The key figure in sequence analysis in sociology is Andrew Abbott, who promoted the method tirelessly from the mid 1980s onwards. His contribution is summarised below in section **ANDREW ABBOTT'S CONTRIBUTION**. While the concern with sequence predates his work, his advocacy of the optimal matching algorithm (see section **The Optimal Matching Algorithm**) has been very influential. A "first wave" of applications can be attributed directly to his influence, which are summarised in a debate in *Sociological Methods and Research* in 2000 (see sections **Arguing with Levine and Wu** for the debate, and **"FIRST WAVE" APPLICATIONS** for the applications). In the years since, there has been a broadening and deepening of the use of sequence analysis, summarised in section **"SECOND WAVE" APPLICATIONS AND FUTURE DIRECTIONS**.

1.2 The Optimal Matching Algorithm

Sequence analysis in sociology is currently almost coterminous with the optimal matching algorithm: there are alternatives (see section **ALTERNATIVE APPROACHES AND SOFTWARE**) but the bulk of research employs this measure. The idea is relatively simple: let the difference between two sequences identical except for one element be related to the difference in the elements, and the difference between two sequences identical except that one has one extra element be the cost of deleting the superuous element; then the difference between any two sequences is the "cost" of the "cheapest" concatenation of substitutions (where sequences differ in elements) and insertions or deletions that change one sequence into the other. For instance, AABC can be changed into ABBD by deleting the initial A, matching on the next A and the B (two zero-cost "substitutions"), inserting a second B, and substituting the C with a D – depending on the relative costing of the operations this may or may not be the cheapest set of operations; for longer sequences there will be many possible sets to consider. The optimal matching algorithm is "optimal" in efciently identifying the cost of this cheapest set of operations. The net result is that OMA can identify similarity (identity or partial similarity, thanks to the substitution operation) at the same or different locations (thanks to insertions and deletions, which permit "alignment", sliding one sequence along another). See section **ALGORITHMIC ORIGINS** for more on the origins of the OM algorithm.

2 ANDREW ABBOTT'S CONTRIBUTION

The American sociologist, Andrew Abbott, has done more than anyone else to introduce sequence analysis to the sociological repertoire, with a long series of publications with colleagues and students, advocating and demonstrating its utility. He has used an eclectic range of applications, and has made strong arguments about the role of an event-focused "narrative" approache to sociology. At times, his advocacy of sequence analysis developed into a trenchant critique of the state of empirical sociology.

2.1 General Evangelism and Applications

Abbott's contribution to the popularization of sequence analysis in sociology has been most effective in presenting real applications in a variety of domains, usually in collaboration with colleagues and students. This process began in the mid-1980s and ran through to the millenium with some vigour, and an eclectic choice of applications.

While Abbott and Forrest 1986 represents the first appearance of OM in sociology, the first paper with real impact was Abbott and Hrycak 1990, on careers of German Baroque musicians. Abbott 1991 addresses trajectories of professionalisation. Abbott and DeViney 1992 takes the welfare state as the unit of observation. Abbott 1995b surveys the empirical and conceptual use of sequence analysis in a broad sense in the social sciences up to that date.

Abbott and Barman 1997 treats the rhetorical organisation of journal articles as a sequential structure.

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16(3).

This is the earliest example of optimal matching in sociology: the sequences are notations of traditional English dances, and the research focuses on patterns of cultural diffusion and change across the different villages where the dances had been recorded.

Abbott, A. and Hrycak, A. (1990). Measuring resemblance in sequence data: an optimal matching analysis of musicians' careers. *American Journal of Sociology*, 96(1).

The first optimal matching paper with real impact. Data on 300 German baroque musicians' careers, looking for evidence of a vacancy-chain process. Coleman clearly erred in telling Abbott: 'No one's gonna pay any attention . . . as long as you write about dead German musicians' (Abbott 2001, p. 13, cited under **Overturning Sociology**).

Abbott, A. and DeViney, S. (1992). The welfare state as transnational event: evidence from sequences of policy adoption. *Social Science History*, 16(2).

An application in yet another domain, the developmental path of the welfare state, demonstrating Abbott's broad notion of what constitutes sequence.

Abbott, A. (1995b). Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, 21.

A survey of sequence analysis in a broad sense, including other disciplines such as psychology and anthropology, and using sequence as a theoretical more than a technical concept. Puts OM in sociology in context, and serves as a good bridge to the papers in the **Overturning Sociology** section.

Abbott, A. and Barman, E. (1997). Sequence comparison via alignment and Gibbs sampling. *Sociological Methodology*, 27

Yet another domain, the rhetorical structure of journal articles considered as sequences. The paper also introduces Gibbs sampling as a solution to some technical problems.

Abbott, A. (1991). The order of professionalization: an empirical analysis. *Work and Occupations*, 18(4).

An application rooted in Abbott's original research interests, professionalisation of occupations viewed as sequences. Theoretically and historically rich.

2.2 Overturning Sociology

Abbott's contribution has two main facets, the promotion of innovative technical methods for dealing with sequence data (outlined under **General Evangelism and Applications**), and a much broader argument about the role of time and sequence in sociological theory. The main elements of his theoretical argument are that the role of time is not handled well in contemporary sociology and that sequence (in the developmental as well as the temporal sense) needs to be taken more seriously. In this he draws on work of David Heise and Peter Abell (Abell

1993; Heise 1989, inter alia) that uses sequence as a tool for theory construction. His theoretical argument evolves from a general view of the importance of sequence in historical sociology (Abbott 1983 and Abbott 1984), takes in a trenchant critique of contemporary sociology as a prisoner of its powerful statistical techniques (Abbott 1988, as "General Linear Reality"), and makes strong claims that sociology needs to move "from units to context, from attributes to connections, from causes to events" (Abbott 1995b, p.93, cited under **General Evangelism and Applications**). This is consonant with the then current critique of "variable-centred sociology" which features in Abell's work, inter alia. At a technical/methodological level this argument has resonances with the notion in Raftery 2001 of the emergence of a "third generation" of statistical methods in sociology in the 1980s, dealing with complex phenomena such as social networks, and longitudinal and spatial data. There is a constant tension in Abbott's work between the ambition of the theoretical argument (Abbott 1990b, Abbott 2001, Abbott 1992) and the practical success of the technical innovations he introduced, the scope of which is much narrower.

Abell, P. (1993). Some aspects of narrative method. *Journal of Mathematical Sociology*, 18, 93–134.

Promotes the concept of narrative method as a theory-building tool, in an approach that developed into a critique of "variable-centred sociology".

Heise, D. (1989). Modeling event structures. *Journal of Mathematical Sociology*, 14, 139–69.

Develops a framework for modelling sociological and historical accounts as event sequences.

Raftery, A. (2001). Statistics in sociology, 1950–2000: a vignette. *Sociological Methodology*, 33

Describes three generations of statisticial methods in sociology, with the third generation focusing on complex phenomena such as social networks, and longitudinal and spatial data. While not explicitly included, sequence analysis clearly fits in this third category.

Abbott, A. (1983). Sequences of social events: concepts and methods for the analysis of order in social processes. *Historical Methods*, 16(4); Abbott, A. (1984). Event sequence and event duration: colligation and measurement. *Historical Methods*, 17(4).

These two papers contain scene-setting arguments, outlining Abbott's views on the importance of sequence as a theoretical tool in historical sociology

Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, 6(2).

In this, Abbott argues that contemporary sociology has become a slave of its strongest tools, and that sequence analysis provides an escape.

Abbott, A. (1990b). Conceptions of time and events in social science methods. *Historical Methods*, 23(4);

A paper in which Abbott develops his argument about the importance of time and sequence in explanation, with reference to history.

Abbott, A. (1992). From causes to events: notes on narrative positivism. *Sociological Methods and Research*, 20(4).

A key statement of Abbott's argument about the importance of narrative, time and sequence in explanation, and the need for a change in sociological argument.

Abbott, A. (2001). *Time matters: on theory and method.* Chicago: University of Chicago Press.

This book brings together a number of useful pieces on time, sequence and sociology.

2.3 Arguing with Levine and Wu

In 2000, *Sociological Methods and Research* published a set of articles on sequence analysis, starting with Abbott and Tsay 2000 who reviewed the progress of optimal matching analysis in sociology. In response Levine 2000 and Wu 2000 raised a series of objections: in particular, they found that Abbott's larger claims about a method that transcends the limitations of sociology were overblown, that OMA could not prove its claims, that no mould-breaking application had been seen, and critically that the sociological meaning of OM distances was fundamentally unclear. While some of the criticisms suggest an incomplete understanding of the algorithm, the critics' repeated point – that findings from a method whose workings are opaque cannot be relied upon – is clear and points to a widely-felt difficulty: how to parameterise OM, and how to interpret the operation of the algorithm, in sociological terms. In biology, similarity of DNA may be interpreted as a measure of distance to a common ancestor, where the steps by which divergence arises are roughly comparable to OM's "elementary operations" of insertion, deletion and substitution, but it is not clear a priori that such operations make sense for social science phenomena. At any rate, Levine and Wu made it clear that sequence analysts needed to do more work to relate distance measures to sociological theories. Abbott 2000 responds to these critiques.

Abbott, A. and Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. *Sociological Methods and Research*, 29(1).

Abbott and Tsay's statement of the state of the art in sociological sequence analysis.

Wu, L. L. (2000). Some comments on "Sequence analysis and optimal matching methods in sociology: review and prospect". *Sociological Methods and Research*, 29(1).

Wu's response.

Levine, J. H. (2000). But what have you done for us lately? Commentary on Abbott and Tsay. *Sociological Methods and Research*, 29(1).

Levine's response.

Abbott, A. (2000). Reply to Levine and Wu. *Sociological Methods and Research*, 29(1).

Abbott's response to their critiques.

3 ALGORITHMIC ORIGINS

The origin of the OMA method is in the Soviet Union, in the so-called Levenshtein distance (Levenshtein 1966). The intended application of Levenshtein's measure was correction of error-prone transmission of information, and the distance itself is a simple count of the number of edits needed to change one sequence into another. "Information transmission" is a very broad concept, and the applications are many, including inexact text search as well as communications. As molecular biology became more computational over the years, techniques derived from the Levenshtein distance, such as OMA, came to be utilised for searching for patterns in macromolecules such as proteins and DNA. In the 1970s Needleman and Wunsch generalised the Levenshtein distance to include the idea of weighted substitutions, and published the algorithm now bearing their name, but also known as the optimal matching algorithm, which can calculate it efficiently (Needleman and Wunsch 1970). Abbott happened across the OM algorithm thanks to meeting Joseph Kruskal, the famous statistician (see the prologue to Abbott 2001 for an account, cited under **General Evangelism and Applications**), at a time when Kruskal and Sankoff were editing a book building on the Needleman and Wunsch algorithm and some other techniques, on methods for analysis of sequence data. This book, *Time warps, string edits and macromolecules* (Sankoff and Kruskal 1983), contains much that has been influential, notably the chapters of Kruskal 1983 (introducing sequence comparison), Bradley and Bradley 1983 (comparing birdsong as sequences), and Kruskal and Liberman 1983 (exploring timewarping as a means of sequence comparison).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady*, 10(8).

Levenshtein's definition of the distance between two token strings in terms of the number of edits (insertions, deletions, substitutions, transpositions) required to change one string into the other, is the starting point for a large body of string comparison algorithms, including optimal matching.

Needleman, S. et al., Wunsch, C. et al. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3).

Adapts the Levenshtein distance to use weighted substitution costs. Describes an efficient algorithm to arrive at minimum-cost alignment between sequence pairs: the "optimal matching" algorithm. Their application was molecular biology, and this marks an early milestone in what is now an extremely large field of bioinformatic sequence analysis.

Sankoff, D. and Kruskal, J. B. (Eds.). (1983). *Time warps, string edits and macromolecules.* Reading, MA: Addison-Wesley. Sankoff and

Kruskal's 1983 book (since re-issued) stands as a major computer science reference for sequence analysis, describing a range of algorithms and applications, and inspiring a substantial amount of further work, primarily in computer science and bioinformatics.

Kruskal, J. B. (1983). An overview of sequence comparison. In D. Sankoff and J. B. Kruskal (Eds.), *Time warps, string edits and macromolecules* (Chap.

1). Reading, MA: Addison-Wesley.

Chapter 1 of Sankoff and Kruskal (1983) explains the concept of alignment and compares it with other methods for sequence comparison, and gives a broad range of applications.

Bradley, D. W. and Bradley, R. A. (1983). Application of sequence comparison to the study of bird songs. D. Sankoff and J. B. Kruskal (Eds.), *Time warps, string edits and macromolecules* (Chap. 6). Reading, MA: Addison-Wesley.

Chapter 6 of Sankoff and Kruskal (1983) shows quite clearly that sequences are not just macro-molecules, with birdsong as the application.

Kruskal, J. B. and Liberman, M. (1983). The symmetric time-warping problem. D. Sankoff and J. B. Kruskal (Eds.), *Time warps, string edits and macromolecules* Reading, MA: Addison-Wesley.

Abbott uses the term "time-warping" in his papers in a loose sense, but here Kruskal and Liberman use it in the more formal sense of comparing sequences by locally compressing and expanding the time axis.

4 DIDACTIC PIECES

A number of authors, from Abbott 1990a onwards, have written papers and books with explicitly instructional intention. Chan 1995 drew on his analysis of Hong Kong data. Abbott returned to the topic in MacIndoe and Abbott 2004. Aisenbrey 2000, Brüderl and Scherer 2004 and Scherer and Brüderl 2010 brought it to a German audience, who also benefitted from the software of Rohwer 1998 and Rohwer and Pötter 2005, and of Brzinsky-Fay et al. 2006 (cited under **Software**). Martin and Wiggins 2011 provide the most recent introduction.

Abbott, A. (1990a). A primer on sequence methods. *Organization Science*, 1(4).

A general discussion of issues relating to analysing sequence data, distinguishing between sequences of recurrent compared with non-recurrent events, and advocating optimal matching for the former.

Chan, T. W. (1995). Optimal matching analysis: a methodological note on studying career mobility. *Work and Occupations*, 22(4).

Somewhat didactic in intent, demonstrating OMA on a small sample of careers using data from Hong Kong.

Aisenbrey, Silke (2000). *Optimal Matching Analyse, Anwendungen in den Sozialwissenschaften* (Optimal Matching Analysis, Social Science Applications). Opladen: Leske+Budrich.

A textbook-style introduction to optimal matching analysis.

MacIndoe, H. and Abbott, A. (2004). Sequence analysis and optimal matching techniques for social science data. M. Hardy and A. Bryman (Eds.), *Handbook of data analysis* Thousand Oaks, CA: Sage.

Another clear didactic advocacy of sequence analysis using optimal matching.

Brüderl, J. and Scherer, S. (2004). Methoden zur analyse von sequenzdata. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 44; Scherer, S.

and Brüderl, J. (2010). Sequenzdatenanalyse. C. Wolf and H. Best (Eds.), *Handbuch der sozialwissenschaftlichen datenanalyse* Wiesbaden: VS Verlag.

Two more works that have been influential in the promotion of sequence analysis in the German-speaking academic world. Rohwer's TDA software (Rohwer 1998; Rohwer and Pötter 2005) and the Stata programs of Brzinsky-Fay et al. 2006, all German-based, have also been influential.

Martin, P. and Wiggins, R. D. (2011). Optimal matching analysis. W. P. Vogt and M. Williams (Eds.). *The Sage Handbook of Innovation in Social Research Methods.* Sage Publications Limited.

A more up-to-date didactic treatment of sequence analysis, covering optimal matching but also Elzinga's combinatorial approaches.

5 "FIRST WAVE" APPLICATIONS

More or less directly inspired by Abbott's advocacy and example, there was a small but significant uptake of optimal matching in sociology through the 1990s. This was predominantly, but not exclusively, US-based, and is well summarised by Abbott and Tsay 2000, cited under **Arguing with Levine and Wu**. The substance covered a range of topics, though with a strong bias to work-life career data: for instance, work careers (Chan 1995, cited under **Didactic Pieces**, and Scherer, 2001); careers of clients of mental health services (Wuerker 1996); transition in careers in Lloyds Bank in the nineteenth century (Stovel, Savage et al. 1996); careers of women in finance (Blair-Loy 1999); temporal patterns of retirement (Han and Moen 1999a); similarity of careers across couples (Han and Moen 1999b); class careers of British and Irish Men (Halpin and Chan 1998); and the temporal pattern of lynching in the southern US (Stovel 2001).

Stovel, K., Savage, M. and Bearman, P. (1996). Ascription into achievement. *American Journal of Sociology*, 102(2).

A sequence-oriented analysis of career data from a British bank, showing a transition between a status-based and an achievement-based system, from 1890 to 1970.

Wuerker, A. (1996). The changing careers of patients with chronic mental illness: a study of sequential patterns in mental health service utilization. *The Journal of Behavioral Health Services and Research*, 23(4).

Treats sequences of services interactions of mental health patients in Los Angeles. A small data set, but of interest because it uses a relatively uncommon form of trajectory.

Halpin, B. and Chan, T. W. (1998). Class careers as sequences: an optimal matching analysis of work-life histories. *European Sociological Review*, 14(2).

Analyses class careers of British and Irish men to age 35 using retrospective data.

Blair-Loy, M. (1999). Career patterns of executive women in finance: an optimal matching analysis. *American Journal of Sociology*, 104(5).

Studies women's career in the finance industry, and identifies change across cohort in opportunity and perspective.

Han, S.-K. and Moen, P. (1999a). Clocking out: temporal patterning of retirement. *American Journal of Sociology*, 105(1).

Looks at trajectories into retirement, noting the ‘multiplex nature of the temporal structuring of the life course’.

Han, S.-K. and Moen, P. (1999b). Work and family over time: a life course approach. *Annals of the American Academy of Political and Social Science*, 562

Examines the extent to which life and work trajectories of couples are coordinated. A relatively rare alternative to cluster analyis of all pairwise distances, in that it uses OM to generate a measure of intra-couple similarity.

Scherer, S. (2001). Early career patterns: a comparison of Great Britain and West Germany. *European Sociological Review*, 17(2).

The school-to-work transition motivates a high proportion of the literature. Scherer uses West German and British panel data, to conduct one of the first OM analyses in a domain where more conventional methods, such as event history analysis or analyses of outcome at a fixed age, have dominated.

Stovel, K. (2001). Local sequential patterns: the structure of lynching in the Deep South, 1882–1930. *Social Forces*, 79(3).

Another alternative to the life-course dominance of sociological sequence analysis, this paper looks at county-level histories of lynching in the Southern US, drawing strongly on arguments from Abbott and others about the necessity of taking a sequence perspective on historical explanations.

6 “SECOND WAVE” APPLICATIONS AND FUTURE DIRECTIONS

Since 2000, there has been vigorous development in the area. Aisenbrey and Fasang 2010 give a very good account of developments in the following decade, arguing that there has been a “second wave” of sequence analysis, and that sequence analysis has moved on despite the problems pointed out by Levine 2000 and Wu 2000 (cited under **Arguing with Levine and Wu**).

A lot of this growth has been in the sociology of the life-course, with emphasis on the transition to adulthood (sections **Life Course: School To Work And The Transition To Adulthood** and **Lifecourse: Other Labor Market Trajectories**), but other sorts of sequence feature too, such as daily time-use patterns (section **Time Use**), residential, activist and organisational careers (section **Some less common types of trajectories**).

Other methods have been proposed (see section **ALTERNATIVE APPROACHES AND SOFTWARE**), as modifications of the optimal matching algorithm or as radical alternatives to it, such as Lesnard’s “dynamic Hamming” measure (with application to time-use data) and Elzinga’s combinatorial methods (which define similarity in terms of sequences going through the same states in the same order, if not consecutively). Simulations and other investigations mean we know a lot more about the characteristics of optimal matching and how to parameterise it.

In parallel, much better software, statistical and graphical tools have emerged, making sequence analysis accessible within widely used statistical packages. While sequence analysis remains predominantly exploratory and descriptive, the hitherto impermeable boundaries between OM and conventional stochastic statistical techniques are being broken, and a large toolbox for dealing with longitudinal sociological data is being built up. We may not observe the revolution against "General Linear Reality" that Abbott advocated, but we have a much better capacity for dealing with the longitudinal in sociology.

Aisenbrey, S. and Fasang, A. E. (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).

An important statement of developments in and applications of sequence analysis in sociology in the first decade of the twenty-first century.

6.1 Life Course: School To Work And The Transition To Adulthood

There has been a substantial take-up of sequence analysis in life-course studies, covering fertility and partnership formation, residence, and the labour market. There is a particularly strong representation in research on the school-to-work transition.

Schoon et al 2001 examines the school to work transition for two UK cohorts. McVicar and Anyadike-Danes 2002 does similar work with a Northern Irish cohort. Brzinsky-Fay 2007 examines labour market entry for cohorts from the European Community Household Panel. Elzinga and Liefbroer address destandardisation of labour market insertion in 19 countries. Martin et al 2008 develop ideal types of labour market entry trajectories. Bras et al 2010 use historical Dutch data on the transition to adulthood. Bühlmann 2010 examines routes into professional and higher managerial work in the UK. Biemann et al 2011 examines destabilization of careers in Germany.

Schoon, I., McCulloch, A., Joshi, H., Wiggins, R. and Bynner, J. (2001). Transitions from school to work in a changing social context. *Young*, 9(1).

Working with data from the UK National Child Development Study and the British Cohort Study (cohorts born 1958 and 1970, respectively), they apply OMA to labour market trajectories of two cohorts, finding a great deal of commonality but a greater difficulty of integration for the younger cohort.

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

Using data from Northern Ireland, the authors use sequence analysis to cluster school-to-work transition trajectories, with a view to identifying factors that predict problematic paths.

Anyadike-Danes, M. & McVicar, D. (2005). You'll never walk alone: childhood influences and male career path clusters. Labour Economics, 12(4), 511–530.

Using data from the 1970 British Cohort Study, the paper identifies early correlates of male trajectories characterised by persistent unemployment.

Elzinga, Cees H., and Aart C. Liefbroer. (2007) "De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis." *European Journal of Population/Revue européenne de Démographie* 23.3-4: 225-250.

Using Elzinga's combinatorial approach to sequence comparison (Elzinga 2005) (see section **Elzinga's combinatorial approaches**), rather than OMA, this paper examines the issue of destandardisation of young adults careers in 19 countries.

Brzinsky-Fay, C. (2007). Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review*, 23(4).

Drawing on data from the European Community Household Panel, this paper addresses the school-to-work transition in ten EU countries, creating a classification that is compared with theoretically-driven classification.

Martin, P., Schoon, I. and Ross, A. (2008). Beyond transitions: applying optimal matching analysis to life course research. *International Journal of Social Research Methodology*, 11(3).

Part of the same team as Wiggins et al. (2007) (cited in section **Life course: multiple domains**), who use theoretically derived ideal types, this paper relates sequences to data-derived ideal types. Distance from each sequence to each ideal type is used to assess the coherence of the classification exercise.

Bras, H., Liefbroer, A. C. and Elzinga, C. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4).

Using combinatorial measures rather than OM (like Elzinga and Liefbroer 2007), this paper tests hypotheses concerning standardisation of pathways to adulthood, with Dutch cohort data.

Bühlmann, F. (2010). Routes into the British service class: feeder logics according to gender and occupational groups. *Sociology*, 44(2).

With data from the UK National Child Development Study birth cohort, this paper examines routes into the salariat, Goldthorpe's "service class". It finds two routes, one direct and one long and "tortuous".

Biemann, T., Fasang, A. and Grunow, D. (2011). Do economic globalization and industry growth destabilize careers? An analysis of career complexity and career patterns over time. *Organization Studies*, 32(12).

This paper focuses on the effect of globalisation and economic change on early careers in Germany. While change is evident, it does not seem to be driven by globalisation, and there is a good deal of stability over time of career patterns.

Liefbroer, Aart C. and Elzinga, Cees H. (2012) Intergenerational Transmission of Behavioral Patterns: How Similar are Parents' and Children's Demographic Trajectories. *Advances in Life Course Research* 17.1: 1-10.

Compares parents' and childrens' lifecourse patterns, and shows that substantial transmission of patterns exists, despite large change in conditions across generations.

6.2 Lifecourse: Other Labor Market Trajectories

While the transition from school to work, or the transition to adulthood in general is a very common focus in the sequence analysis literature, there is also work that examines other parts of the labour market career. Malo and Muñoz-Bullón 2003 and Levy et al 2006 examine labour market careers over a longer perspective, and Fasang 2012 focuses on the transition to retirement.

Malo, M. A. and Muñoz-Bullón, F. (2003). Employment status mobility from a life-cycle perspective: a sequence analysis of work-histories in the BHPS. *Demographic Research*, 9

Using retrospective work-life history data from the British Household Panel Study, the authors use OMA to deepen a descriptive view of careers in the mid-twentieth century.

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

Data from the Swiss Household Panel Study is used in this paper to address the question of the extent to which either standardisation or individuation of occupational careers is observed, by gender and across cohort.

Fasang, Anette E (2012): Retirement Patterns and Income Inequality. *Social Forces* 90(3): 685-711.

This paper uses sequence analysis to make sense of changing patterns of retirement in Germany and the UK.

6.3 Lifecourse: Multiple Domains

In many ways the exploratory potential of sequence analysis is highest where the sequences are complex, and the inter-relation of multiple domains (such as work, education, family and fertility) is a special case of complexity. Quite a bit of work has been done using multiple domains, either by combining domains into a single complex state space, or by using "multi-channel sequence analysis" (MCSA) software (for instance Gauthier et al 2010).

A focus on multiple domains is present early on: Dijkstra and Taris 1995 (cited under **Early alternatives**) used multiple domains in their example, and Han and Moen 1999a (cited under **"FIRST WAVE" APPLICATIONS**) also used multiple work-life domains. Aasave et al 2007 combines labour market, partnership formation and fertility of British women. Pollock 2007 also uses British data, on housing, marital status and children. Wiggins et al 2007 look at the effect of work, housing and partnership histories on the elderly. Bühlmann 2008 combines occupation and industry classifications. Müller et al 2008 combine residence, partnership and fertility for Swiss data. Gauthier et al 2010 make a formal argument for multi-channel sequence analysis, and present examples using specialised software.

Aasave, A., Billari, F. and Piccarreta, R. (2007). Strings of adulthood: analyzing work-family trajectories using sequence analysis. *European Journal of Population*, 23(3-4).

One of a small group of papers that pushes sequence analysis in the interesting direction analysing multiple life-course domains simultaneously, “multichannel sequence analysis”. Using females’ life histories from the British Household Panel Study, it unites the labour market, partnership formation and childbearing, to create a data-driven typology.

Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A*, 170(1).

A nice example of a multi-channel sequence analysis bringing together employment, housing tenure, marital status and responsibility for children, using ten years of British Household Panel Study data.

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

Predicting quality of life in old age from housing, partnership and work life-histories, this paper uses optimal matching to assign trajectories to ideal types developed a priori, an effective but relatively unusual approach in the literature. See also Martin, Schoon et al. (2008).

Bühlmann, F. (2008). The corrosion of career? Occupational trajectories of business economists and engineers in Switzerland. *European Sociological Review*, 24(5).

An explicitly multi-channel sequence analysis paper, taking account simultaneously of different dimensions of the work-career of engineers and economists in Switzerland.

Müller, N. S., Lespinats, S., Ritschard, G., Studer, M. and Gabadinho, A. (2008). Visualisation et classication des parcours de vie. *Revue des Nouvelles Technologies de l’Information*, 2(E-11).

Uses retrospective data from the Swiss Household Panel to examine multidimensional life-course trajectories (residence, partnership, fertility), arguing for the utility of OM for exploratory analysis of longitudinal data. Uses multidimensional scaling to reduce the complexity of the distance data, and proposes useful visualisation techniques for trajectory data.

Gauthier, J.-A., Widmer, E. D., Bucher, P. and Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology*, 40(1).

Formally argues for multi-channel sequence analysis rather than parallel single-channel: calculating distances based on all dimensions together produces more robust and informative results than calculating distances on each dimension separately. Uses custom software.

6.4 Some less common types of trajectories

While a great deal of sequence research focuses on labour market and family issues, a certain amount of research on other life-course issues is evident. Clark et al 2003 and Stovel and Bolan 2004 focus on different aspects of housing careers, Blanchard 2010 examines the careers of political activists, and Stark and Vedres 2006 uses OM to model the ownership trajectories of firms.

Clark, W. A. V., Deurloo, M. C. and Dieleman, F. (2003). Housing careers in the United States, 1968-93: modelling the sequencing of housing states. *Urban Studies*, 40(1)

This paper explores housing trajectories, with a focus on the form of tenure.

Stovel, K. and Bolan, M. (2004). Residential trajectories: using optimal alignment to reveal the structure of residential mobility. *Sociological Methods and Research*, 32(4).

Again a paper on housing, but with a focus on location.

Blanchard, P. (2010). Analyse séquentielle et carrières militantes. Rapport de recherche

This paper examines the engagement careers of political activists.

Stark, D. and Vedres, B. (2006). Social times of network spaces: network sequences and foreign investment in Hungary. *American Journal of Sociology*, 111(5).

Here the longitudinal subject is the firm, in particular the ownership trajectories of Hungarian firms.

6.5 Time Use

Time use research is a very distinct area within sequence analysis, with Laurent Lesnard's "dynamic Hamming" distance measure as the dominant alternative to OMA. The dynamic Hamming measure compares sequences time-point by time-point (no "alignment") but using inter-state distances dynamically calculated from the time-dependent pattern of rates of transition between the states. Thus at times when lots of transitions are occurring, states are judged to be more similar to each other than at times when the transition rate is low. Lesnard makes a strong argument that this is more appropriate for time-use data when the "clock" or calendar is important, and the dislocation of time caused by alignment is inappropriate. Lesnard 2010 makes the argument for the appropriateness of dynamic Hamming as an alternative to OM. De Saint Pol 2007 focuses on the distinctive timing of meals in France. Lesnard 2008 uses dynamic Hamming distance on time date for dual-earner couples. Lesnard and de Saint Pol 2009 uses dynamic Hamming to examine the scheduling of paid work.

de Saint Pol, T. (2007). Le dîner des français : un synchronisme alimentaire qui se maintient. *Économie et Statistique*, 400

This paper uses conventional OMA to analyse the timing of meals in French time-diary data, showing the persistence of a very distinct pattern of cultural agreement about when to eat.

Lesnard, L. (2008). Off-scheduling within dual-earner couples: an unequal and negative externality for family time. *American Journal of Sociology*, 114(2).

This paper uses dynamic Hamming to focus on the effect of employment on family time for dual-earner couples.

Lesnard, L. and de Saint Pol, T. (2009). Patterns of workweek schedules in France. *Social Indicators Research*, 93

Another paper using dynamic Hamming, to look at the scheduling of paid work time.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research*, 38(3).

Here Lesnard outlines in detail the argument for his dynamic Hamming distance measure.

7 ALTERNATIVE APPROACHES AND SOFTWARE

The optimal matching algorithm is by some margin the dominant approach in sequence analysis in sociology. In this section some alternatives are presented. The two most prominent alternatives are dynamic Hamming (see section **Time use**) and Elzinga's combinatorial approaches (see section **Elzinga's combinatorial approaches**). Sequence analysis prior to the availability of OM is discussed under **Early Alternatives**, systematic attempts to assess OM under **Evaluating OM and Its Parameterisation**, and two attempts to improve on OM under **Directly Modifying the OM Algorithm**. Other papers discussing alternatives to OM are discussed under **Competing and Complementary Approaches** and model-based approaches under **Competing Approaches: Model Based**. Software issues are briefly discussed under **Software**.

7.1 Early Alternatives

Researchers were applying other methods to the problem of the holistic examination of complete sequences before OM became available. The simplest approach is to define similarity on the basis of element-by-element similarity (similarity at the same time), so-called "Hamming distance". While this cannot pick up similarity displaced in time in the way that OMA can, it has the virtue of clarity. Buchmann and Sacchi 1995 used Hamming distance in a relatively sophisticated manner, using collateral data on occupations to define distances in a very large state space (i.e., hundreds of occupational groups) and then classifying work-life careers of a Swiss birth-cohort. In this much it is clear that OMA was not a solution in search of a problem, but served a need for simplifying complex longitudinal data that was already felt by researchers.

Another approach to determining similarities between sequences was to divide them into periods (e.g., combine monthly data into 6-month sections) and conduct a factor analysis on summaries (such as cumulated durations) of the periods (Degenne et al. 1996). This allowed a certain amount of time-dislocation (within the blocks), while retaining a computationally simple period-by-period Hamming-style comparison between the blocks, using distances based on the factors. This is sometimes known as qualitative harmonic analysis (QHA, see also Robette and Thibauld 2008, cited under **Evaluating OM and Its Parameterization**) for a recent comparison of this method and OMA).

Another competing approach was that of Dijkstra and Taris 1995, who proposed a method to define similarity or distance, which involves dropping elements not shared by both sequences, and counting the number of common subsequences (a focus on the same states in the same order). In a comment on that paper, Abbott could demonstrate that OMA was a much more general measure (Abbott 1995a). Dijkstra and Taris's method (implemented in Macintosh software, see Dijkstra 1994, cited under **Software**) is no longer used, but some of its motivation has been taken up by Elzinga (see **Elzinga's combinatorial approaches**) in his subsequence-oriented approaches.

Buchmann, M. and Sacchi, S. (1995). Mehrdimensionale Klassifikation beruflicher Verlaufsdaten: eine Anwendung auf Berufslaufbahnen zweier Schweizer Geburtskohorten. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 47(3).

This paper precedes optimal matching, but works with a definition of sequence similarity (of work-life careers) in terms of distance (at the same time) between occupations using a sophisticated multi-dimensional approach to defining distance between occupations.

Degenne, A., Lebeaux, M.-O. and Mounier, L. (1996). Typologies d'itinéraires comme instrument d'analyse du marché du travail. Typologie des marchés du travail, suivi et parcours. Document du Céreq, 115.

A rare example of Qualitative Harmonic Analysis (see also Robette and Thibauld 2008 **Evaluating OM and Its Parameterization**), applied to French work-life data.

Dijkstra, W. and Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods and Research*, 24(2).

The paper in which they propose their definition of sequence similarity.

Abbott, A. (1995a). A comment on "Measuring the agreement between sequences". *Sociological Methods and Research*, 24(2).

Abbott's rebuttal, in which he shows that the OM algorithm is more general.

7.2 Evaluating OM and Its Parameterisation

A good deal of the critique of OM focuses on how to set the various parameters, and the consequences of the details of the algorithm for inter-sequence distances (e.g., Wu 2000 cited under **Arguing with Levine and Wu**). Some work that addresses these problems directly includes Wilson 2006, which uses simulation to explore how well sequence analysis can recover underlying structures, Robette and Thibauld 2008 which compares qualitative harmonic analysis (see also Degenne et al 1996, cited under **Early Alternatives**) with OM, and Gauthier et al 2009 which addresses the parameterisation of OM directly.

Wilson, C. (2006). Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning A*, 38(1).

This paper uses a range of techniques with simulated data, to assess how well underlying structure can be recovered, in part in response to Levine (2000).

Robette, N. and Thibauld, N. (2008). Comparing Qualitative Harmonic Analysis and Optimal Matching. *Population*, 63(4).

Compares OM with Qualitative Harmonic Analysis, a technique which compares sequences in terms of summaries of blocks (e.g., six-month spans), using much more straightforward comparison at the cost of losing some (but not all) of the sequential information.

Gauthier, J.-A., Widmer, E. D., Bucher, P. and Notredame, C. (2009). How much does it cost?: optimization of costs in sequence analysis of social science data. *Sociological Methods and Research*, 38(1).

Much of the criticism of OM takes the form of claims that parameterisation is a dark art, and that there is no sociological basis for assigning substitution and insertion and deletion costs (Levine 2000; Wu 2000). This paper is a relatively rare example of analysts addressing the issue of substitution costs directly.

7.3 Directly Modifying the OM Algorithm

OM focuses on editing token strings, with elementary operations that focus on tokens out of context, such that, for instance, substitution takes account only of one token in each string, ignoring the tokens' neighbours. This can be seen as sociologically unattractive, as for instance by Wu 2000, cited under **Arguing with Levine and Wu**. Hollister 2009 and Halpin 2010 independently propose distinct but similar adaptations of the OM algorithm to take account of context. Unfortunately, neither algorithm preserves the metric property of the OM distance.

Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research*, 38(2).

Hollister proposes "localised OM", where the cost of insertion and deletion operations are modied according to the substitution cost between the inserted (deleted) element and its neighbours. Thus changes that make less substantive difference are cheaper.

Halpin, B. (2010). Optimal matching analysis and life course data: the importance of duration. *Sociological Methods and Research*, 38(3).

Halpin's argument is that operations on tokens in long spells should cost less than on tokens in short spells. His duration-adjusted OM adapts the original algorithm to weight operations inversely with the square root of the spell length.

7.4 Competing and Complementary Approaches

A number of papers offer competing or complementary approaches to the use of OM in sequence analysis. Billari 2001 presents a "monothetic divisive" algorithm for non-repeating event data, which has structure that OM will not exploit. Billari et al 2006 presents techniques drawing from the machine learning literature. Piccarreta and Lior 2010 and Piccarreta 2012 focus on multi-dimensional scaling as an alternative to cluster analysis of inter-sequence distances. Studer et al 2011 presents the notion of "discrepancy" which permits ANOVA-like decomposition of pairwise inter-sequence distance matrices.

Billari, F. C. (2001). Sequence analysis in demographic research. *Canadian Studies in Population*, 28(2).

A primer for sequence analysis in general, this paper also presents a "monothetic divisive" algorithm which classifies "unique-event" data (first job, first partnership, first birth, etc) in an efficient and interpretable manner.

Billari, F. C. & Piccarreta, R. (2005). Analyzing demographic life courses through sequence analysis. Mathematical Population Studies, 12(2), 81–106.

Further development of the ideas introduced in Billari 2001.

Billari, F. C., Fürnkranz, J. and Prskawetz, A. (2006). Timing, sequencing and quantum of life course events: a machine learning approach. *European Journal of Population*, 22

Another innovative approach from Billari and colleagues, looking at machine-learning techniques that exploit the timing, order and amount of events. The application is to Italian and Austrian transitions to adulthood.

Piccarreta, R. and Lior, O. (2010). Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society Series A*, 173(1).

Sequence analysts typically run cluster analysis on the pairwise distances, but alternatives are possible. In this paper Piccarreta and Lior argue for the utility of extracting dimensions from the distance matrix using multi-dimensional scaling. This imposes a simple element of order on the sequences, enabling exploratory research.

Piccarreta, R. (2012). Graphical and smoothing techniques for sequence analysis. *Sociological Methods and Research*, 41(2).

This paper extends the MDS approach of Piccarreta and Lior (2010) to dealing with very large data sets, and proposes a means of smoothing the graphical representation, greatly facilitating exploration and description. A downside is that the first MDS dimension will often exclude a lot of interesting variation.

Studer, M., Ritschard, G., Gabadinho, A. and Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40(3).

An alternative to cluster analysis for analysis of pairwise inter-sequence distances: the "discrepancy" measure, of average distance to group centre. Permits ANOVA-like calculations using observed categorical variables, or cluster solutions. Offers pseudo-F tests and a pseudo-R2 measure, using bootstrapping. A very useful complement to cluster analysis or MDS of distances.

7.5 Competing Approaches: Model Based

Optimal matching provides an algorithmic distance measure (which we may use to generate a typology) but model-based distances and model-based classifications are also possible. Massoni et al. 2009a use enhancements of Markov models (characterizing sequences by a relatively parsimonious structure of time-weighted transition parameters), while the other papers use Latent Class models to group sequences in a stochastic manner. Vaughn et al 2009 focus on criminal careers, and Barban and Billari 2012 directly compare sequence analysis with latent class analysis. As sequence analysis matures, it is important that links are being made between the exploratory and algorithmic techniques and more conventional statistical techniques.

Massoni, S., Olteanu, M. and Rousset, P. (2009a). Career-path analysis using drifting Markov models (DMM) and self-organizing maps; Massoni, S., Olteanu, M. and Rousset, P. (2009b). Career-path analysis using optimal matching and self-organizing maps. Proceedings of the 7th international workshop on advances in self-organizing maps (WSOM 2009) (5629). *Lecture Notes in Computer Science*. Berlin/Heidelberg: Springer.

Defines distance between sequences according to the parameters of a "drifting Markov model". Also use Kohonen self-organizing maps as an alternative to conventional cluster analysis

Vaughn, M. G., DeLisi, M., Beaver, K. M. and Howard, M. O. (2009). Multiple murders and criminal careers: a latent class analysis of multiple homicide offenders. *Forensic Science International*, 183

Uses Latent Class Analysis to categorize criminal careers.

Barban, N. and Billari, F. (2012). Classifying life course trajectories: a comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society Series C*, 61(5).

A comparison of Latent Class Analysis and OM, using simulation. While the techniques give largely consistent results, they credit OM with more sensitivity with respect to timing and ordering of events.

7.6 Elzinga's Combinatorial Approaches

In an approach that can be related conceptually to Dijkstra and Taris 1995 (cited under **Early alternatives**), in that it focuses on similarity defined as the extent to which sequences experience "the same states in the same order", Elzinga 2003, Elzinga 2005, and Elzinga 2010 propose a set of methods that count how often sequences share the same subsequences (i.e., order-preserving subsets of the sequences; AC is a subsequence of ABC). This produces a measure of longitudinal similarity that has a radically different justification compared with OM: its focus is on shared order rather than similarity with allowance for time-displacement. Elzinga et al 2008 links Elzinga's developments in social sequence analysis with parallel developments in computer science and biology. Bras et al 2010 apply Elzinga's method to Dutch transition to adulthood data.

Elzinga, C. H. (2003). Sequence similarity: a non-aligning technique. *Sociological Methods and Research*, 32(1).

The original statement of the approach as a superior alternative to OM and Dijkstra and Taris's approach.

Elzinga, C. H. (2005). Combinatorial representations of token sequences. *Journal of Classification*, 22(1).

A more extensive and comprehensive statement of the approach. It presents an efficient combinatorial algorithm to enumerate common subsequences.

Elzinga, C. H., Rahmann, S. and Wang, H. (2008). Algorithms for subsequence combinatorics. *Theoretical Computer Science*, 409(3).

Elzinga's combinatorial algorithms have no counterpart in the social science literature, but Wang and Rahmann independently developed similar ideas in computer science and biological domains, and in this paper they bring their ideas together.

Bras, H., Liefbroer, A. C. and Elzinga, C. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4);

Two applications of Elzinga's method to Dutch data, focussing on the transition to adulthood. See also Elzinga and Liefbroer 2007 (cited under **Life Course: School To Work And The Transition To Adulthood**).

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods and Research*, 38(3).

This paper proposes a measure of complexity of sequences, based on combinatorial arguments. While it is quite technical, the notion of sequence complexity is of great sociological importance.

7.7 Software

Sankoff and Kruskal 1983 (cited under **ALGORITHMIC ORIGINS**) explained the OM algorithm with sufcient clarity that Abbott was able to implement it in BASIC. Later it was implemented in Rohwer's TDA (Rohwer 1998, and Rohwer and Pötter 2005), and is now available in R in the TraMineR package (Gabadinho et al. 2009, 2011) and Stata (see Brzinsky-Fay et al. (2006), and also Halpin's SADI package for Stata). Elzinga's subsequence methods are implemented in his own software, CHESA, and some are also available in TraMineR and SADI. Lesnard's dynamic Hamming is availabe both in TraMineR and SADI. Of historical interest is Dijkstra 1994, which implements the method of Dijkstra and Taris 1995 (cited under **Early alternatives**).

Dijkstra, W. (1994). Sequence – a program for analysing sequential data. *Bulletin de Méthodologie Sociologique*, 43

In this paper Dijkstra presents his "Sequence" program for analysis in the framework of Dijkstra and Taris (1995).

Rohwer, G. (1998). Transition data analysis. Computer program; Rohwer, G. and Pötter, U. (2005). *TDA user's manual.* Universität Bochum.

TDA or "Transition Data Analysis" was a ground-breaking package for the analysis of longitudinal data, initially conceived of as a work-bench for event-history analysis techniques. Rohwer added optimal matching capabilities in the mid-1990s, and for a long time it was the only practical way of conducting sequence analysis.

Brzinsky-Fay, C., Kohler, U. and Luniak, M. (2006). Sequence analysis with Stata. *Stata Journal*, 6(4).

Sequence analysis capabilities were added to the Stata program as a user-written package, as described in this paper. The package has strengths in visualisations, but is somewhat slow. Other possiblities for Stata exist.

Gabadinho, A., Ritschard, G., Studer, M. and Müller, N. S. (2009). *Mining sequence data in R with the TraMineR package: a user's guide for version 1.2.*

University of Geneva.

The TraMineR package for the statistical programming language R, is now the most powerful and general package for sequence analysis.

Gabadinho, A., Ritschard, G., Müller, N. S. and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4).

A presentation of the very powerful TraMineR package for sequence analysis using the R statistical programming language, with an emphasis on its visualisation capabilities.