Rejoinder: Positivism and Big-game Fishing :a reply to comments

Nicolas Robette, Xavier Bry, Eva Lelièvre

Positivism and big game fishing: a reply to comments

Nicolas Robette, Printemps (UVSQ-CNRS, UMR 8085)

Xavier Bry, I3M, Université Montpellier 2

Éva Lelièvre, INED

> "Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise." (Tukey 1962)

The publication of our article in *Sociological Methodology* was the successful conclusion of a long "sequence" involving one journal's refusal to referee it (for not being in their field), a presentation at the RC33 conference of the ISA in 2012, and submission to *SM* followed by six rounds of revisions—a record both for that journal and for us. The symposium concerning the article has thus gone much further than we would have hoped when we started the work in 2009. Indeed, we are grateful for the opportunity we have thus had to exchange with specialists in sequence analysis (and life course analysis) and in this way construct a necessarily partial and temporary situation report on the progress made by this family of techniques. We do not have the space in this rejoinder to discuss all the criticisms and observations we have received. However, certain patterns have emerged and we shall try to address the most "robust". First, we must point out and rectify a few misunderstandings.

It all depends

The first misunderstanding concerns the contrast between local and global interdependence, which we explain in the first part of the article. This contrast as we see it is a "conceptual" one, in the sense that it concerns the way the interdependence between the dimensions of the sequences is "grasped" and recorded by statistical techniques. It therefore precedes the chain of analysis. For example, in order to study the life courses of two spouses after they formed a couple it is appropriate to consider these life courses as being simultaneous (because they develop jointly within each couple) and to compare the couples from point to point. This is a case of local interdependence and MCSA is particularly appropriate. One of us has indeed used MCSA in exactly this sort of case (Pailhé et al. 2013). But if we now turn to the study of homogamy on the basis of the two spouses' past life courses leading to their forming a couple, their alignment from point to point makes little sense and this is a case of global interdependence, to be analyzed with GIMSA.

Fasang, in her commentary, appears to understand this distinction between local and global interdependence in a different sense, concerning the interpretation of results, i.e. subsequent to the analysis chain: are the dimensions of the sequences substantively linked in a general manner or at certain specific points in their development? This recalls the event/sequence dichotomy or, in Billari's terms, that between the atomistic and the holistic approach (Billari 2001).

Under our definition of the distinction between local and global interdependence, the comparison between MCSA and GIMSA is less relevant, since the two techniques do not address the same problem. Which is why in the article we compare GIMSA and Strategy 4,

both of which address global interdependence[1].

The fact that Fasang's comparison of MCSA and GIMSA on the basis of our application leads to similar results does not imply that the two techniques are interchangeable and one might reasonably choose the simpler. This comparison reveals rather the existence of deeper structural patterns in the analyzed data (as is often the case with empirical data in the social sciences).

Inflexibility goes before a fall

The difference between MCSA and GIMSA, therefore, is "conceptual." But it is also practical: GIMSA analyzes dimensions of multidimensional sequences of varying length, with different time windows (age vs. calendar years, for example) and time units (years, months, etc.), and uses different metrics for each dimension so as to emphasize a particular aspect of time (order, duration, date)[2]. MCSA can just about cobble together the data formatting, aligning differing dimensions of length, time windows and time units, by using missing value states, for example (cf. Fasang 2015), but the sociological significance of this "forced" alignment remains questionable[3]. Last but not least, MCSA uses a single metric.

GIMSA's practical flexibility, pointed out by a number of commentators (especially Pavalko 2015; Fan and Moen 2015), is one of the elements in its added value (disputed in other commentaries). It involves a series of choices, seen by some as a weakness, in the sense that the user is no longer perfectly "controlling" what they are doing and the "robustness" of the method is allegedly weakened by this. We cover robustness issues in the following section, but here we shall merely note that this concerns a debate between flexibility and simplicity like that about Optimal Matching some years ago; a debate in which it is not for us to take sides[4].

Each of GIMSA's steps has an unmistakable and indispensable role, with a varying latitude of decision:

- Choosing a dissimilarity measure is indeed necessary when tracking patterns, since these emerge from similarity groups. Can methods that allow only one dissimilarity measure be viewed as more robust because they preclude choice? It all depends on whether this dissimilarity measure is indeed unique for theoretical reasons. Failing that, the fact that GIMSA can support various possible choices of dissimilarity measure should not be viewed as a drawback, but as an asset: it should lead the researcher to justify their choice of one measure over others, or to use several

[1] However, by producing mother and daughter clusterings independently, Strategy 4 cannot find the "channels of information" where the mother-daughter link is strongest, i.e. explore the content of the mother-daughter similarity, which is partial. But it is instructive to look further into these partial similarities restricted to certain "channels of information" (unknown and requiring further research). That is precisely what GIMSA does in its PLS stage.

[2] In our application, the two dimensions are substantively different, because with mothers we are examining their occupational career (in terms of social class) and with daughters their school-to-work transition (in terms of employment). These are of differing lengths, one defined by age and the other by an initial event (i.e. leaving school) and two different metrics are used. Ideally, we would have preferred to have the information in months for the daughters, but we make do with the same time unit for both dimensions (namely the year). But the aim of our paper was to briefly present a new methodology we had built up, not to investigate a sociological issue in depth.

[3] Gauthier (2015) proposes an alternative, namely placing the dimensions in a single sequence end to end. There too, the substantive meaning of this formatting is dubious. Furthermore, this approach involves using a single alphabet for all dimensions.

[4] Although we tend to the view that flexibility is a virtue, both intellectually and methodologically.

measures and compare their outputs, looking for discrepancies as well as invariants across them;

- Multidimensional Scaling involves no real decision. It just translates the dissimilarity into the closest Euclidean distance, and outputs the corresponding coordinates of units;
- Canonical PLS searches both spaces for principal "directions of matching." This also involves no choice other than the number of retained directions. This choice is a necessary compromise between the richness of the description of the matching (in terms of dimensions) and its quality: the more dimensions we retain, the less strong the matching. Now, any method concerned with matching should have the following two concerns: 1) providing the ability to tune the demanded level of matching quality; 2) keeping the dimensions of noise (i.e. dimensions carrying structurally weak information) away from those considered in the matching. PLS is one of the simplest ways to achieve that, since it involves no tuning-parameter[5]. Any regularized type of canonical correlation analysis could also be used here[6], on condition the regularization is based on the structural strength of the components so as not to find correlations between noisy (i.e. non-information-bearing) features. This is the value of PLS;
- The clustering step seems to us the one that involves questionable choices. It is also the only non-compulsory step in GIMSA: after identifying the structural "dimensions of matching" (previous step), one could analyze them in terms of life-history events by correlating them with all kinds of life-history descriptors, so without having to perform clustering. Clustering is rightly famous for the many arbitrary choices it demands. This echoes the fuzziness of its root-question: "what is similar to what, how, and in what respect?" But here, the final clustering is but one of the many ways to interpret the dimensions of matching. Ideally, these dimensions should be analyzed in a number of alternative ways, in order to extract the maximum amount of the information they capture.

GIMSA's flexibility means that it is a particularly suitable instrument for studying linked lives, but, like sequence analysis in general, its potential field of application goes beyond life course analysis. Consequently, we would invite colleagues to disinhibit their "sociological imagination" as Mills (1959) recommended, and include data that are perhaps richer than they habitually use.

Guilty by association?

The second misunderstanding concerns the aims of GIMSA and the analysis of multidimensional sequences generally. As Studer astutely points out in his comment, the cluster analyses we habitually use are not designed to analyze the degree of association between dimensions and are not suited to do so. We obviously agree with this: GIMSA, like MCSA, is a pattern search technique, no more, no less. We plead guilty to sloppy use of vocabulary (noted by Studer), particularly in the description of the results of the application,

[5] The PLS components may admittedly lack certain forms of association, but that is true of any methodology. PLS looks for linear correlations between strong (less noisy) dimensions and finds them. Furthermore, non-linear extensions of this PLS stage may perfectly well be envisaged via the Reproducing Kernel Hilbert Space technology.

[6] In the test carried out by Piccarreta (2015), canonical correlation analysis provides almost the same results as PLS because the "denoising" has been done previously, and not all the MDS components are used. But it may be preferable to keep all the MDS components and apply PLS to them.

where we tended to overuse “link” terminology. The clustering step cannot, and therefore should not, be interpreted as a way of finding connections, but rather as a way of broadly summarizing the connections teased out by the PLS components submitted to clustering. This clustering step is only secondary anyway: GIMSA is mainly the combination of the first three steps (see above). Here too, we make no claim to be doing any more than fishing for patterns of dyads of sequences[7]. This remark may disarm some of the criticisms made of GIMSA, for it can easily be seen that they are indeed expressed in terms of the degree of association between dimensions.

What is “pattern searching” in social sciences about?

This misunderstanding evokes more serious differences of opinion about how to envisage the use of statistics in social sciences. When Andrew Abbott introduced Optimal Matching into the world of social sciences in the 1980s, this took its place within a broader discussion of what he calls “general linear reality” (Abbott 2001a; Robette 2015). He sees the “methodological framework” of the social sciences as being structured by a set of dichotomies: quantitative versus qualitative, positivism versus interpretation, etc. (Abbott 2001b, p.28). These dichotomies possess “elective affinities,” of which the most profound associates positivism with analysis and narrative with interpretation. Abbott seeks to break down these affinities by reintroducing a narrative dimension into positivism. This means proposing an alternative to the “paradigm of variables” that dominates quantitative empiricism, and its implicit presuppositions (Abbott 2001a; Fabiani 2003). The analysis of sequences provides a set of tools for developing this alternative, among which Abbott singled out OM. In 2000, an article by Abbott and Tsay in *Sociological Methods & Research* was followed by two comments by Levine and Wu. Levine takes up a firm position in favor of general linear reality, reproaching OM mainly for not meeting the standards of stochastic models. Wu, a specialist in event history analysis, adopts the same point of view and also formulates more targeted criticisms of particular aspects of the method, such as the sociological meaning of the operations of substitution, insertion and deletion of elements within sequences, and the inclusion of the order of the events in the sequences. In response to all these criticisms, Abbott corrects what he sees as miscomprehensions about the workings of OM, and more particularly resituates the method within the dichotomy of general linear reality versus narrative-descriptive methods: any assessment of OM against the bases of mainstream statistical methods *de facto* invalidates most of the criticisms (Abbott 2000):

> OM algorithms are not models, nor are they premised on models. That is the foundation of their difference from standard methodologies. They simply look for patterns or regularities. The type of regularity they seek can be varied by varying the structure and parameters of the algorithm. But the algorithms do not rest, ultimately, on an idea of how the data are generated.

And yet in this symposium, just as more broadly in the assessments of research on the basis of sequence analyses, the criticisms have often been founded on principles close to criteria of scientificness calqued on those of the experimental sciences, i.e. on an “instrumental positivism” as defined by Bryant, who calls it “‘instrumental’ insofar as it is the available research instruments that mark out the object of research, and ‘positivist’ because this self-imposed constraint of sociologists reflects their desire to submit to an analytical rigor similar to that they attribute to the natural sciences” (Bryant 1989 [retranslated]).

For example, in his comment, Elzinga considers that a degree of agreement of 0.65 between two clusterings is not satisfactory, contrary to what we state in our article. In his view, one

[7] Which also implies that we make no causal hypothesis about possible links between dimensions (such as “Dimension A causes Dimension B”). That is why we use symmetric PLS rather than asymmetric PLS.

cannot settle for a value below 0.9. He illustrates this with some amusing and revealing examples: the allocation of children to one educational program or another, and of patients to one therapy or another. But that is precisely the point; we are not policy makers or doctors but social scientists: decision making is far beyond our scope. In quantitative sociology research it is common practice to use a significance threshold of 5%. This is merely a statistical habit: who would undergo vision correction surgery if medical engineering tolerated a similar degree of error? Many commonly accepted rules for statistical choices in our disciplines are social constructs, traditions based on no real theoretical foundations. These choices can only be contextual and often empirical, and any normative aspiration is founded on a poor understanding of the particular epistemology of the social sciences (Passeron 1991). The general problem of thresholds is easy to understand: just try to answer the question, "how many grains of sand make a sandpile?"

This "instrumental positivism" recurs in the matter of the number of classes of typology produced by sequence analysis. Again and again, the referees of articles we have submitted to various journals (and here *Sociological Methodology* is no exception) came back with remarks like "there is no 'numerical' or 'statistical' criterion mentioned to motivate the choice of a cluster solution." Lurking in the background is the idea that there is a "true" solution, or at least a "best" solution, which statistical tools are intended to reveal.

However, any automatic classification procedure will place all the individuals in a study population into mutually exclusive groups. So any of the possible solutions is "true." As for which is the "best," no general answer can be given, even for a single set of data: it all depends on the research question, the interpretability of the results and their value for advancing current sociological themes, the use to be made of the typology, etc[8]. As Williams and Lance (1965), cited in our article, assert, a typology is not true or false, it is profitable or unprofitable. They add, "To define an optimum method we should have to formalize the situation sufficiently to estimate, and thence to maximize, the expected profitability. The purpose of such methods is not to displace the intuitive taxonomist, but to suggest to him potentially fruitful lines of investigation." To base the choice of number of classes on a statistical criterion is less a guarantee of scientificness on the researcher's part than an abdication of responsibility.

But our view does not appear to be widely shared: as Aisenbrey and Fasang (2010) note, the "validation" of sequence analysis results is repeatedly criticized. They suggest a remedy might be to use cutoff criteria based on the dispersion of within- and between-cluster distances and take the best solution to be the number of classes at the point where the ratio of within- to between-cluster distances falls below 0.5 for the first time [9]. But what is the theoretical basis for this threshold? It is merely a heuristic. Furthermore, there are many cutoff criteria, and they do not necessarily lead to the same conclusions: so it is easy for the cunning researchers to choose the criterion that suits them best so as to satisfy their peers while preserving their own choices. The whole apparatus of validity tests, robustness checks, sensitivity tests and "noise models" may well have some use, but mainly for improving one's

[8] A 24-cluster solution may be instructive at an exploratory stage but will turn out to be hard to reconcile with the characterization of classes by logistic regression and the summary presentation of results for a scientific journal article. Furthermore, it may be useful to remember that good practice advises closely studying various partitions of the same classification.

[9] They add that another validity criterion is met when the groupings found with sequence analysis relate to variables as theoretically expected. This reflects a use of sequence analysis restricted to validating hypotheses, and therefore not open to the possibility of "discovery." We think an exploratory stage ("fishing for patterns") is yet necessary before comparing patterns with any "well-established theory," in order for the currently admitted theory to be given a chance to evolve under the pressure of observations.

chances of being published in the leading journals by aping the experimental sciences' criteria of scientificness. The wisest thing to do when taking an exploratory, heuristic and non-confirmatory approach would be to 1) use as many instruments as possible that are apparently technically suited to identifying the patterns one wishes to discover (e.g. correlations, partitions), with a wide range of values for their tuning parameters; 2) compile and critically interpret the similarities and differences between the results obtained, so as to sort out the more robust patterns from the weaker ones (those depending most on the observation instrument), or even from pure artifacts via meta-analysis.

When Benzécri developed correspondence analysis in 1962-1965, he was hoping to "discover the hidden properties, higher in the natural hierarchy of causes than those that are obvious, which control the obvious ones" (Benzécri 1973). In his view, therefore, "since the realities of this world are things created by God, the statistician's work is to work back from the facts to the essence of things, the shape the Creator gave them," (Cibois 1981). Those using this technique immediately set aside these philosophical foundations. But one may well wonder whether, driven out by the door, these ideas have not slipped back in through the open window of mainstream statistics in its quest for the "true" or the "best" solution[10].

With correspondence analysis, Benzécri also intended to introduce into France a way of doing and seeing statistics similar to the data analysis practiced by English-speaking researchers (Cibois 1981), which one may describe as follows, "It designates not really a set of techniques, let alone an 'established doctrine,' but rather 'a certain idea of statistics' whereby it is legitimate in principle (even if in practice problems arise) to examine the data in order to interpret them, whatever the intentions and procedures of their collection may have been, without the need to confine oneself to a model or restrictive hypotheses," (Rouanet and Lépine 1976). It is within this legacy, we believe, that pattern search techniques such as sequence analysis should be placed.

"What are you going to do for us presently?"

Once these misunderstandings have been cleared up, we may attempt now to summarize the encouraging prospects for research into sequence analysis outlined by the comments in this symposium.

First, as has been argued, the automatic classification of multidimensional sequences is not a tool for examining the degree of association between dimensions. However, the question of the association between dimensions is a central one and there are already some ideas for research in that direction. Elzinga suggests using distance matrices of the various dimensions, analyzing their association from Mantel, Kendall or R_v coefficients and another coefficient based on the notion of "local monotonicity" (see also Piccarreta and Elzinga 2013). Studer mentions Cramer's *V* and standardized Pearson residuals (to analyze the contingency table of typologies for each dimension), discrepancy analysis (see also Studer et al. 2011) and "sequences of typical states" based on implicative statistics (see also Studer 2012). Taken together, these techniques already provide a copious toolbox, which we should use and test more widely.

Nearly thirty years after OM was introduced into the social sciences, the question of comparing metrics remains open. A number of studies of systematic comparison have shown that many existing metrics gave closely similar results, although some metrics do stand out (Robette and Bry 2012; Studer and Ritschard 2014). Indeed, the recent Subsequence Vector Representation metrics (SVR) seem particularly effective when focusing on the order of

[10] How much our scholarly practices and habits of thought in the social sciences owe to this deep, longstanding infusion of experimental science epistemology in our university courses, handbooks, editorial boards, etc.—even among those of us who attempt to deny the fact—is worthy of a study in itself.

elements within sequences (Elzinga and Wang 2013; Elzinga and Studer 2015). We should bear in mind that the choice of metric, although it certainly does not fundamentally alter the results, is no trivial matter, and it may be instructive to test a number of metrics on one set of data before proceeding with analyses[11].

Piccarreta's point is also important: "can sequences be so easily substituted by the MDS scores?" Abbott and De Viney appear to say yes in their article on policy adoption sequences (1992)[12]. MDS applied to the matrix of distances between national sequences enables them to identify two main structuring factors, interpreted as the timing of pensions program adoption and the timing of health insurance adoption. These two factors are then analyzed separately as dependent variables. However, MDS only provides a Euclidean approximation of a dissimilarity that is not necessarily Euclidean. Any Euclidean metric is perfectly rendered by the full set of MDS components[13], whereas a non-Euclidean metric is only rendered approximately. So the question is, what information is lost by substituting MDS components for the distance matrix originally chosen, i.e. what is the "non-Euclidean share" of this distance? Thorough research would be needed into ways of finding the Euclidean within the non-Euclidean. Use of MDS for sequence analysis probably deserves wider investigation (Piccarreta and Lior 2010) before we adopt it as a matter of routine.

Finally, one last prospect for research is the connection between the local and global (here in Fasang's sense), i.e. event and sequence. As Fan and Moen point out in their comment, one might, for example, ask "how a given transition in one person's life is tied to temporal patterns in another's." The path toward combining the standard tools of event history analysis and those of sequence analysis appears at first blush to be a stony one in both technical and epistemological[14] terms, but it may not be totally impassable. Studer's (2012) "sequences of typical states" may well supply another line of enquiry. Let us bet that this is the direction that will be taken by the most stimulating innovations in sequence analysis in the years ahead.

References

Abbott, Andrew. 2000. "Reply to Levine and Wu." *Sociological Methods and Research* 29:65-76.

Abbott, Andrew. 2001a. *Time matters. On theory and method*. Chicago : The University of Chicago Press.

Abbott, Andrew. 2001b. *Chaos of disciplines*. Chicago : The University of Chicago Press.

Abbott, Andrew and Stanley De Viney. 1992. "The Welfare State as Transnational Event: Evidence from Sequences of Policy Adoption." *Social Science History* 16(2):245-274.

Abbott, Andrew and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology." *Sociological Methods and Research* 29:3-33.

Aisenbrey, Silke and Anette 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* 38:420-62.

[11] At this step, as at others, it is more appropriate to choose on the basis of the researcher's own intelligence than of statistical validation criteria (see above).

[12] See also Halpin and Chan 1998.

[13] MDS is a noise reduction method only if not all the components it provides are retained. This is therefore a secondary property and a relatively accessory one.

[14] Causality vs pattern searching, inference vs description, etc.

Benzécri, Jean-Paul. 1973. *L'analyse des données*. Paris : Dunod.

Billari, Francesco. 2001. “Sequence analysis in demographic research.” *Canadian Studies in Population* 28(2):439-458.

Bryant, Christopher. 1989. “Le positivisme instrumental dans la sociologie américaine.” *Actes de la Recherche en Sciences Sociales* 78:64-74.

Cibois, Philippe. 1981. “Analyse des données et sociologie.” *L'Année Sociologique* 31:333-348.

Elzinga, Cees H. 2015. “On the association between sequences: a comment on ‘Global Interdependent Multidimensional Sequence Analysis’.” *Sociological Methodology*.

Elzinga, Cees H. and Matthias Studer. 2015. “Spell Sequences, State Proximities, and Distance Metrics.” *Sociological Methods and Research* 44(1):3-47.

Elzinga, Cees H. and Hui Wang. 2013. “Versatile string kernels” *Theoretical Computer Science* 495:50-65.

Fabiani, Jean-Louis. 2003. “Pour en finir avec la réalité unilinéaire. Le parcours méthodologique de Andrew Abbott.” *Annales. Histoire, Sciences Sociales* 58(3):549-565.

Fan, Wen and Phyllis Moen. 2015. “Comment on ‘Globally Interdependent Multiple Sequence Analysis’.” *Sociological Methodology*.

Fasang, Anette E. 2015. “What’s the Added Value? Commentary on ‘GIMSA, a new approach to uncover patterns of linked life courses’.” *Sociological Methodology*.

Gauthier, Jacques-Antoine. 2015. “Comments on "A 'Globally Interdependent Multiple Sequence Analysis' approach to uncover patterns of linked life courses".” *Sociological Methodology*.

Halpin, Brendan and Tak Wing Chan. 1998. “Class careers as sequences: an optimal matching analysis of work-life histories.” *European Sociological Review* 14(2):111-130.

Levine, Joel H. 2000. “But what have you done for us lately? Commentary on Abbott and Tsay.” *Sociological Methods and Research* 29(1):34-40.

Mills, Charles W. 1959. *The Sociological Imagination*. Oxford University Press.

Pailhé, Ariane, Nicolas Robette, and Anne Solaz. 2013. “Work and family over the life course. A typology of French long-lasting couples using optimal matching.” *Longitudinal and Life Course Studies* 4(3):196-217.

Passeron, Jean-Claude. 1991. *Le Raisonnement sociologique. L'espace non-poppérien du raisonnement naturel.* Paris : Nathan.

Pavalko, Eliza K. 2015. “Bridging the Gap Between Life Course Concepts and Methods.” *Sociological Methodology*.

Piccarreta, Raffaella. 2015. “Simplifying sequences using scores: Some considerations.” *Sociological Methodology*.

Piccarreta, Raffaella and Cees H. Elzinga. 2013. “Mining for Associations between Life Course Domains.” In *Contemporary Issues in Exploratory Data Mining*, edited by J.J. McArdle and G. Ritschard, Quantitative Methodology Series, chapter 8. Routledge : New York.

Piccarreta, Raffaella and Orna Lior. 2010. “Exploring sequences: a graphical tool based on multi-dimensional scaling.” *Journal of the Royal Statistical Society Series A* 173(1):165-184.

Robette, Nicolas. 2015. “Du prosélytisme à la sécularisation. Le processus de diffusion de l’‘Optimal Matching Analysis’.” in *Andrew Abbott, sociologue de Chicago*, edited by M. Jouvenet and D. Demazière. Paris: Editions de l'EHESS. *Forthcoming*.

Robette, Nicolas and Xavier Bry. 2012. “Harpoon or Bait? A Comparison of Various Metrics in Fishing for Sequence Patterns.” *Bulletin of Sociological Methodology* 116:5-24.

Rouanet, Henry and Dominique Lépine. 1976. “A propos de ‘l'Analyse des données’ selon Benzécri : Présentation et commentaires.” *L'année psychologique* 76(1):133-144.

Studer, Matthias. 2012. *Étude des inégalités de genre en début de carrière académique à l’aide de méthodes innovatrices d’analyse de données séquentielles*. Thèse de doctorat n°777, Faculté des sciences économiques et sociales, Université de Genève.

Studer, Matthias. 2015. “On the use of globally interdependent multiple sequence analysis.” *Sociological Methodology*.

Studer, Matthias and Gilbert Ritschard. 2014. “A comparative review of sequence dissimilarity measures.” *LIVES Working Papers* 33, NCCR LIVES, Switzerland.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas Müller. 2011. “Discrepancy analysis of state sequences.” *Sociological Methods and Research* 40(3) :471-510.

Tukey, John W. 1962. “The Future of Data Analysis.” *The Annals of Mathematical Statistics* 33(1):1-67.

Williams, W.T. and G.N. Lance. 1965. “Logic of computer-based intrinsic classifications.” *Nature* 207(4993):159-161.

Wu, Lawrence L. 2000. “Some Comments on ‘Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect.’” *Sociological Methods and Research* 29:41-64.