*Journal of Interdisciplinary History*, XVI:3 (Winter 1986), 471–494.

*Andrew Abbott and John Forrest*

# Optimal Matching Methods for Historical Sequences

> In some rural riots, as in the French [riots] of 1789, the path of disturbance actually followed well-trodden and traditional routes.[1]

The notion that riots and other collective disturbances follow a common script is standard historical and sociological fare. In the passage from which this quote is drawn, Rudé describes a script proceeding from general grievances to triggering events and on to a fixed repertory of rick-burning, machine-breaking, and so on. Throughout his earlier chapters he presents stories of individual riots, on the evidence of which this generalized sequence rests. But one may well ask when we can really speak of such a fixed sequence of events. How would we identify such a sequence quantitatively? How could we discover if the characteristic sequence in rural riots differed sharply from that of urban disturbances? How might we analyze the effects of the occupations of participants on such sequences of events?

The answers to such questions are not easily found. The problem of event sequence in riots is typical of a much larger category of sequence problems in history. In each case the data are lists of events. In each case the analyst wants to separate the common pattern from its particular realizations. Yet in each case the analyst wants also to understand why each particular realization took the form it did. The only practical approach to these tasks has been to generate a common pattern using an "ideal type" or comparative analysis, and then to consider the variations from it on an individual basis. There have been no effective quantitative methods for analyzing such "sequence data," a fact that goes far

Andrew Abbott is Assistant Professor of Sociology at Rutgers University. John Forrest is Assistant Professor of Anthropology at S.U.N.Y., Purchase.



1 George Rudé, *The Crowd in History* (New York, 1964), 242.

toward explaining why narrative sequences are considered to be exclusively the province of traditional history.[2]

But a peculiar scientific conjuncture has recently created methods that will answer those quantitative questions. In this research note we discuss these methods and apply them to a classic historical problem: Has a certain sequence of events changed its character over time? In our case the sequence involved is nothing so complicated as Rudé's riots. It is the sequence in which figures are performed in a particular set of ritual dances. This example has the virtue of simplicity. There is a simple theoretical issue. There is a simple data set. There is nothing to obscure the sequence problem itself. But, although so self-contained an example illustrates the methods with particular clarity, it does not show the breadth of their applicability. Therefore, we close the note with a short discussion of how one might apply the methods to solve more complex problems.

THE RISE OF OPTIMAL MATCHING TECHNIQUES Sequence data have long been a problem for statisticians. Yet their fundamental principles are straightforward. A sequence is simply an ordered listing of items, which may be events, numbers, or anything else. In some sequence data sets, each item appears once and only once in each sequence; in others each item may appear several times in some sequences and not at all in others. In the first case, the problem is to analyze permutations: in the second, it is to analyze recurrent events.

Sequences of recurrent events have proved much less tractable than permutations. Among the probability models applied to them, only the Markov model has had wide application in social science. But Markov models cannot deal with the kinds of questions that interest historians. How do we find a typical se-

2 Most attempts to explain riots and revolutions have focused on structural factors. E.g., Theda Skocpol, *States and Social Revolutions* (Cambridge, 1979); Jeffery M. Paige, *Agrarian Revolution* (New York, 1975). This is true even of Rudé's own work, e.g., *The Crowd* and (with Eric J. Hobsbawm) *Captain Swing* (New York, 1968). For overviews of the "compare by cases" method, see Skocpol and Margaret Somers, "The Uses of Comparative History in Macrosocial Inquiry," *Comparative Studies in Society and History*, XXII (1980), 174–197; Arthur Stinchcombe, *Theoretical Methods in Social History* (New York, 1978), 89–97. An overview of the sequence problem is Abbott, "Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes," *Historical Methods*, XVI (1983), 129–147.

quence? How do we analyze differences between several different versions of "the classic riot"?[3]

Fortunately, problems of this kind have become more pressing in a number of other sciences. In computer science, the problem arose of comparing long sequential files for errors and omissions. If the tenth of 10,000 characters was omitted from one of two exactly similar files, then a pairwise comparison of the files would reveal uniform disagreement (except for random matches caused by repeated symbols) after the tenth item. Yet the files would in fact be extremely alike, differing only in this one omission. A human would notice the similarity at once; a computer comparing pairs of symbols would not. The string edit problem, as it was called, became a central focus for authors of file-comparison algorithms.[4]

Similar problems arose in the comparison of proteins and other large molecules. Analytical methods enabled the mapping of immense molecules as long strings of a small number of elements—amino acids in the case of proteins, nucleotides in the case of DNA. Once these huge maps were available, biochemists wished to see whether some parts of these molecules repeated or very nearly repeated other parts. Again methods were necessary for measuring how similar two sequences were to one another. Even more important, sequence methods could be used to create evolutionary "trees" of molecules, on the assumption that similar molecules were related by some sort of descent. Sequential analysis of macromolecules thus offered major insights into biochemical evolution.

Another area of research that employed sequence comparison was speech processing. Again the problem was similar. A given

3 Space considerations have led us to cut most comparative references to the standard methods for permutation data. Complete permutation data support multilinear permutation statistics that generalize the Spearman rank correlation coefficient. With the censored data and real time characteristic of real historical data, the required techniques are the so-called seriation methods. The best general discussions of these methods are F. Roy Hodson, David G. Kendall, and Petre Tăutu (eds.), *Mathematics in the Archeological and Historical Sciences* (Edinburgh, 1971); J. Douglas Carroll and Phipps Arabie, "Multidimensional Scaling," *Annual Review of Psychology*, XXXI (1980), 607–649. The original article on this subject was W. S. Robinson, "A Method for Chronologically Ordering Archeological Deposits," *American Antiquity*, XVI (1951), 293–301.

4 The basic review of the relevant literature is David Sankoff and Joseph B. Kruskal (eds.), *Time Warps, String Edits, and Macromolecules* (Reading, Mass., 1983). The following general discussion of optimal matching techniques is based largely on Kruskal's opening chapter in that book, "An Overview of Sequence Comparison," 1–44.

phrase sounds different when spoken by different speakers, and methods were sought that would recognize a phrase despite these variations. Each version of a phrase is a sequence of sounds, and analysts sought a method that would recognize different versions as a cluster of sequences with a single written referent. The problem was the more difficult in that speech sequences exist in continuous time, not in discrete space, as do macromolecules and computer strings.

Methods designed to solve these and other related questions arose simultaneously in a number of fields. The methods consist of two parts: first, an operational concept of the distance between two sequences, and second, a method for finding the route that minimizes that distance. Often called optimal matching methods, these methods are directly applicable to sequence problems in history. The typical historical sequence problem is the one we outlined above. One wants to find a characteristic sequence, or several characteristic sequences, among a large sample of sequences. One seeks to explain variation among these sequences in terms of some external variables. To illustrate how optimal matching methods solve such a historical problem we analyze a case involving Cotswold morris dances.

The empirical problem examined here is part of a larger study using Cotswold morris dances to trace patterns of solidarity in rural England during the nineteenth century. Traditional theory among students of these dances has been that each village did the dances in its own way and that the patterns remained remarkably constant over time. That theory seems surprising given historians' views of rural upheaval in England, and we sought to test it rigorously. Fortunately, one crucial part of the village way of doing the dances is the sequence of certain figures. Hence we can use the stability or instability of the figure sequence as an indicator of the stability or instability of the village dance tradition, and, by implication, of the way of life of which these traditions were a part. For us optimal matching methods present a direct way of testing an operational hypothesis about the English countryside.

DANCE SEQUENCES AND VILLAGE TRADITIONS The morris dances, with their colorful costumes, ringing bells, and clashing sticks, were collected in detail by Sharp and others during the period from 1880 to 1920. Morris dancing had flourished from about

1750 to 1860. It began to decline around 1850, and would have been lost altogether but for the industry of the collectors and revivalists after 1880. The morris dancers of a village were collectively known as the village's side or team. Many Cotswold villages, and a few villages elsewhere in the surrounding area, had morris sides. From the early days of the morris revival, it was assumed that each village side embodied a tradition. The Ascott-under-Wychwood dancers waved their hands thus and so, whereas Bampton-in-the-Bush did so and thus.[5]

The existence of strong village traditions in morris dance would be a matter of considerable theoretical interest. The agricultural community in England was rapidly changing in the late eighteenth century. Yet Sharp imagined that morris dances were ancient traditions based on fertility rituals and uniquely characteristic of each village. The morris villages typically contained less than 800 people, many of whom were born and bred elsewhere. Yet these tiny groups, without benefit of written records, were thought to have maintained separate cultural traditions that were stable across decades and even centuries. To be sure, the Cotswolds were a conservative area agriculturally, late to come to full enclosure, late to take up the scythe and other agricultural innovations, and, in the morris villages at least, seemingly immune to the Swing agitation of the 1830s. But secure dance traditions maintaining their independence and purity at distances that were often less than five miles would show such astonishing stability that we would be forced to rethink English rural history

5 The collectors on whose work we have drawn are the following: D'Arcy Ferris, mss. in the Vaughan Williams Memorial Library, London; Percy Manning (with Tom Carter), mss. in the Bodleian Library, Oxford; Cecil James Sharp, mss. at Clare College, Cambridge; *idem* and Herbert C. McIlwaine (pts. I–III) and George Butterworth (pt. V), *The Morris Book* (London, 1907, 1909, 1910, 1911, 1913); Sharp and McIlwaine, *The Morris Book* (London, 1912, 1919, 1924; 2d ed.). The dates 1750–1860 are somewhat speculative, especially the earlier. For a full account of the relevant historical material see Keith Chandler, "A Biographical Index to Traditional Dancers, Fools, and Musicians," *Morris Matters*, II (1979), 3:8–17, and 4:26–33; *ibid.*, III (1980), 1:20–27, and 2:24–35; E. C. Cawte, Alex Helm, R. J. Marriott, and Norman Peacock, "A Geographical Index of the Ceremonial Dance in Great Britain," *Journal of the English Folk Dance and Song Society*, IX (1960), 1–41; Forrest, *Morris and Matachin: A Study in Comparative Choreography* (Sheffield, 1984). On the existence of village traditions, see *idem*, "Here We Come A-Fossiling," *Dance Research Journal*, forthcoming. The acceptance of village tradition is shown in Lionel Bacon's classic work, *A Handbook of Morris Dances* (London, 1974). All 300 dances described therein are classified by village tradition.

in this period. It was therefore important to decide whether village traditions in the dance really did exist.[6]

The foundation of Sharp's concept of village tradition was the sequence of dance figures or patterns. There are five basic types of figures in morris dancing, and numerous versions of each one. Morris dances typically proceed by a sequence in which these figures alternate with choruses. According to Sharp, choruses tended to be common to a dance wherever it was danced. What was unique to the *village* and *common* from dance to dance within a village was the sequence of figures proper.

These sequences of dance figures exemplify historical data on sequences of potentially recurrent events. To recall our opening example, the dances unfold like the "contentious gatherings" of the same period, studied by Tilly, Rudé, Hobsbawm, and others. They proceed from figure to figure on to their end, much as a riot goes from "arrive" to "cheer" to "attack" to "disperse." The same questions are relevant to each: Are there characteristic sequences? Do these change with time? Even the substantive questions are similar: Is there a traditional way to riot, or a traditional way to dance? If we had data on figure sequences at several points in time, but within a single town, we could see whether there was any variety, both between dances at a given time, and between the times taken as units. Optimal matching methods could then be used to measure such variety, and thereby give the stability of village tradition a direct test.[7]

For one town at least, over-time data on dance sequences are available. There are descriptions of the dances in Ilmington, Warwickshire, for four distinct periods. The dances were directly

6 On agricultural England during this period, see Hobsbawm and Rudé, *Captain Swing*; Pamela Horn, *Labouring Life in the Victorian Countryside* (Dublin, 1976). Village population and mobility data were obtained from the relevant volumes of *Victoria County History*. The conservative qualities of Cotswold agriculture are noted in E. J. T. Collins, "Harvest Technology and Labor Supply in Britain 1790–1870," *Economic History Review*, XXII (1969), 453–473; M. K. Ashby, *Joseph Ashby of Tysoe* (Cambridge, 1961); A. F. Martin and R. W. Steel, *The Oxford Region* (Oxford, 1954). Hobsbawm and Rudé's lists of Swing riots contain only one town with known morris dance—Noke, on Otmoor. See also Bernard Reaney, *The Class Struggle in Nineteenth-Century Oxfordshire* (Oxford, 1970).

7 Charles Tilly and R. A. Schweitzer, "Enumeration and Coding of Contentious Gatherings in Nineteenth-Century Britain," unpub. ms. (Univ. of Michigan, 1980). An elegant analysis suggesting that there was in fact a traditional way to riot is Edward P. Thompson, "The Moral Economy of the English Crowd in the Eighteenth Century," *Past & Present*, 50 (1971), 76–136.

observed in two periods, from 1887 to 1897 and from 1906 to 1910. The first period preceded the serious morris revival; the latter occurred when the revival had just begun. These two observed dance periods are supplemented by constructed data on two other periods. For the year 1867, Ilmington's dances were reconstructed by Sharp, who felt that the 1887 dances were mutually inconsistent; he sought to restore the purity of the old side, which had lapsed in the earlier year. There are also constructed dance data for 1945, which have a more complex history. Sharp's victory in various controversies among revivalists had a distinct effect on Sam Bennett, the musician and founder of the 1906 Ilmington side. After 1910, Bennett tried to appease Sharp by revising his dances to conform with Sharp's ideas in *The Morris Book*. The dances perpetuating this reconstruction were observed by Schofield in 1945 and subsequently published by Bacon. There are, then, four waves or cohorts of data. The first, 1867, and the last, 1945, contain constructed data. In the first instance, no dances were observed; in the second, the dances observed were based on published sources concerning the first. The two middle waves, however, reflect observations of unconstructed dances and interviews with the Ilmington men who performed them.[8]

The Ilmington dances use only a limited number of the Cotswold morris figures. There are approximately seventy-five identifiable figures, of which Ilmington dancers use twenty-two. These twenty-two are drawn from the five basic types of figures: once-to-yourself, footing, partners, rounds, and hey. In Ilmington dancing, there are two different once-to-yourself's, six different footings, nine partners, three rounds, and two heys. The actual sequences of the dances are shown in Table 1, organized by dance and cohort. We have used notation for four dances: "Shepherd's Hey," "Black Joke," "Maid of the Mill," and "Bumpus o'Stretton." Each dance is shown in as many different versions as were either observed or constructed. Thus, since the 1867 version of "Shepherd's Hey" could be closed with either of two versions of

8 Our data are as follows: (1) the 1867 dances (dances "of the old side") as published by Sharp in *The Morris Book*; (2) the 1887–1897 dances from Sharp's manuscript notes; (3) the 1906–1910 Bennett dances from Bacon, *Handbook,* and from the manuscripts of Mary Neal, at the Vaughan Williams Memorial Library in London; (4) the 1945 dances from R. Kenworthy Schofield's manuscript notes, at the Vaughan Williams Memorial Library in London, and from Bacon, *Handbook*.

*Table 1* Ilmington Dance Sequences

| SHEPHERD'S HEY | | BLACK JOKE | |
|---|---|---|---|
| 1867A | O1,F5,P1,P4,P8,R1 | 1867A | O2,F5,P1,P4,P8,R1 |
| 1867B | O1,F5,P1,P4,P8,R2 | 1867B | O2,F5,P1,P4,P8,R2 |
| 1887A | O1,F5,P4,P3,R2 | 1887A | O2,F2,R1 |
| 1887B | O1,F5,P4,P1,R2 | 1906A | O2,F6,P7,F4,F6 |
| 1906A | O1,F3,P5,F4,P5,F3 | 1945A | O2,F5,P2,P9,R2 |
| 1906B | O1,F3,P7,F4,P5,F3 | | |
| 1906C | O1,F3,P5,F4,P7,F3 | | |
| 1906D | O1,F3,P7,F4,P7,F3 | | |
| 1945A | O1,F5,P2,P9,R2 | | |

| MAID OF THE MILL | | BUMPUS O'STRETTON | |
|---|---|---|---|
| 1867A | O1,F5,P1,P4,P8,R1 | 1867A | O1,F5,P1,P4,P8,R1 |
| 1867B | O1,F5,P1,P4,P8,R2 | 1867B | O1,F5,P1,P4,P8,R2 |
| 1887A | O1,F5,F1,P6,F1,P6,H1,F1,P6,F1,P6,R1 | 1887A | O1,F5,P3,R1 |
| 1906A | O1,F3,P5,F6,H2,P7,F6,P5,F6,F4 | 1906A | O1,F3,P5,F4,P5,F3 |
| 1945A | O2,F5,P9,H2,P9,F6,P9 | 1906B | O1,F3,P7,F4,P5,F3 |
| | | 1906C | O1,F3,P5,F4,P7,F3 |
| | | 1906D | O1,F3,P7,F4,P7,F3 |
| | | 1945A | O1,F5,P2,P9,R2 |

O = Once to Yourself
F = Footing
P = Partners
H = Hey
R = Rounds

rounds, it appears as two different versions of the dance. Because of these variant versions, there are in fact eight dance sequences from the first wave, five from the second, ten from the third, and four from the fourth—a total of twenty-seven dance sequences.

THE OPTIMAL MATCHING METHOD The Ilmington dances, then, give us the over-time data necessary to test our hypothesis about the stability of the dance traditions. Since they are sequences of recurrent events (figures can repeat), optimal matching is the appropriate form of analysis. Optimal matching methods differ from the seriation methods usually applied to permutation data in a central respect; they measure distance between *sequences* rather than distance between *events*.

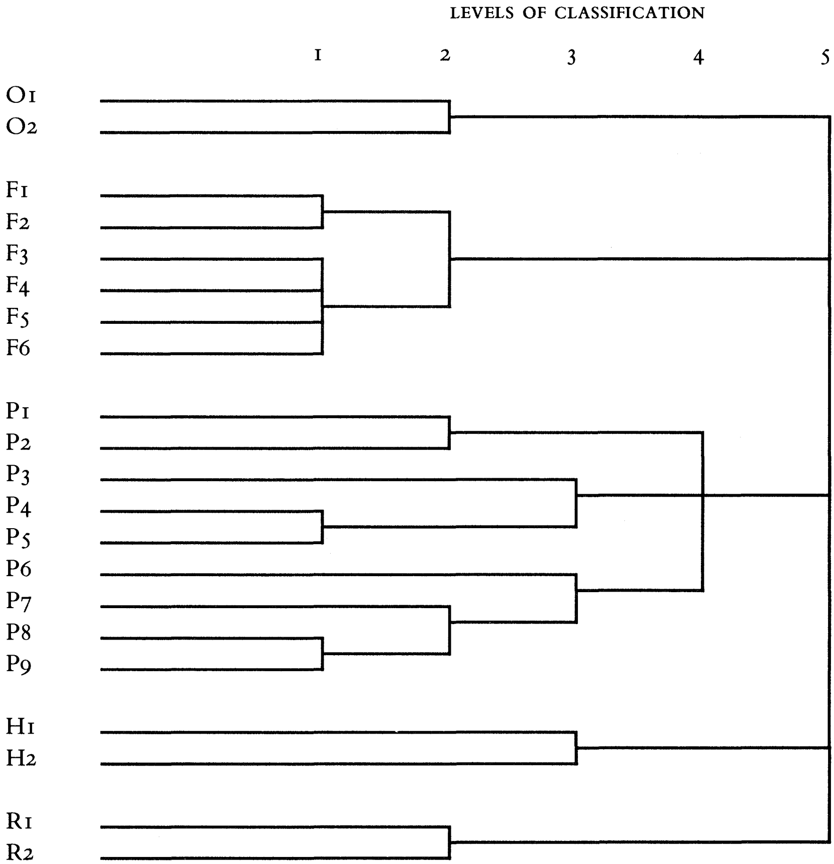
Like DNA sequences, dances can be changed in two basic ways. First, one can replace a figure with another figure. Second, one can insert or delete a figure. In some sense, the "distance" between two dance sequences ought to be a function of the number of these elementary operations—substitutions on the one hand, and insertions and deletions (collectively known as "indels") on the other—required to transform one sequence into another. Since there are many ways to make this transformation, just as there are many ways to go from one place to another on a city block grid, so the distance should be a function of the *minimum* number of these elementary operations that can accomplish the transformation. Finally, some elementary operations are more costly than others. Replacing a footing figure by a rounds figure is clearly more drastic than replacing it by another version of a footing figure. Thus, the minimum we seek should take into account the costs of various substitutions, insertions, and deletions.

The first step in the analysis is to estimate the costs of all possible substitutions and/or indels. This estimation can be accomplished by a variety of means. We have done it by categorizing the figures into a hierarchy. Each pair of figures can be characterized by the number of steps up the hierarchy necessary to put them under a common heading. The ratio of this number to the total number of steps possible, in this case five, is the cost of substitution. The categorization of figures is shown in Table 2. One can see from the table that substituting O2 for O1 has a cost of two over five, or .40. All substitutions that change the type of figure (e.g., substituting F1 for P1) have a cost of 1.0. It is possible, if one desires, to create several such hierarchies and combine the results. Our hierarchy is in fact based on one classification only. Costs of insertions and deletions can also be set by judges, but we have used a uniform value equal to the highest substitution cost (1.0).

The second step in analysis is to calculate the minimum cost paths between each pair of dances, a total of 351 pairs. There are various recursive algorithms for this procedure; these are the optimal matching methods proper. Table 3 illustrates them with dance data. Two dance sequences, those for the *first* versions of "Shepherd's Hey" in 1867 and 1887 respectively, are used.[9]

9 David W. Bradley and Richard A. Bradley, "Application of Sequence Comparison to the Study of Bird Songs," in Sankoff and Kruskal, *Time Warps,* 189–209. This work was particularly helpful because it compared large numbers of fairly short sequences. In most applications, comparison is between two very long sequences, and the problems of application are different. For the analysis reported here, we have used software kindly

*Table 2* Taxonomy of Figures




The algorithm creates an alignment of the two sequences. An alignment is a row of vertical pairs: the upper member of each pair is in one sequence; the lower member is in the other. The procedure begins with a null pairing and adds pairs of various cost until the upper and lower sequences are complete. To simplify matters, we have assumed in this example a uniform cost of 1.0 for each indel or substitution. The one exception is the substitution of a figure for itself, a vertical pair in which the upper

provided by David Bradley, to whom we are greatly indebted. The software includes not only programs for entering sequences, setting costs, and minimizing transformation operations, but also programs for scaling, clustering, and grouping procedures.

*Table 3* Example of Optimal Matching

| | NULL | O1 | | F5 | | P1 | | P4 | | P8 | | R1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NULL | 0 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 4 | 1 | 5 | 1 | 6 |
| O1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 0 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 4 | 1 | 5 |
| F5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 4 |
| P4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 3 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 3 |
| P3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 4 | 1 | 3 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 3 |
| R2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 5 | 1 | 4 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 |

THREE ALIGNMENTS YIELDING SEQUENCES [O1] AND [O1]

| A. | | B. | | C. |
|---|---|---|---|---|
| O1 | NULL | NULL | O1 | O1 |
| NULL | O1 | O1 | NULL | O1 |

FINAL ALIGNMENT

| | | | | | |
|---|---|---|---|---|---|
| O1 | F5 | P1 | P4 | P8 | R1 |
| O1 | F5 | NULL | P4 | P3 | R2 |

and lower elements are identical. This substitution has a cost of zero. In the common sense we are not substituting anything, but rather adding equivalent items. It is, however, convenient to use a single term for adding *any* vertical pairing to the alignment, and to define the substitution of equals as having zero cost.

The algorithm creates the alignment by working across the matrix shown. Each interior cell has four entries. The one in the lower left gives the cost of "adding a deletion," that is, adding a vertical pair to the alignment in which the upper element is (say) O1, while the lower element is null. (Null elements are disregarded in reading the sequences from the alignment.) The entry in the upper right gives the cost of "adding an insertion," that is, adding a pair in which the upper element is null, while the lower element is (say) O1. The upper left element gives the cost of "adding the substitution" of, say, O1 over (for) F5.

There are three ways to move from the null pairing to the exactly equivalent sequences [O1] and [O1]. Alternatively, we could say that there are three alignments from which we can read the identical sequences. The three are shown in Table 3. One involves adding a deletion and then adding an insertion; the second involves adding an insertion and then adding a deletion; and the third involves adding the substitution of O1 for itself. The first two have a total cost of two, whereas the third has a total cost of zero, so that the third is the least cost path. These three ways move through the matrix in three different patterns: the first goes into the (null,O1) cell, then into the (O1,O1) cell; the second goes into the (O1,null) cell, then into the (O1,O1) cell; and the third goes directly from the (null,null) cell to the (O1,O1) cell.

The number in the lower right corner of each cell is the least cost to arrive at that cell. This number may be found in a straightforward manner. Each cell has three predecessors—one above it, one to its left, and one above and to the left. (Any other predecessors would involve moving backwards in the alignment.) Each predecessor has a lower right-hand element giving the least cost of arriving there. From each predecessor one incurs an insertion, deletion, or substitution cost in arriving at the present cell; these are the upper right, lower left, and upper left elements of the cell, as we have just seen. To each predecessor's least cost of arrival is added the relevant cost of entering the present cell. The smallest of the three sums is written in the lower right-hand corner of the present cell and the algorithm goes on to the next. The reader can work across the matrix and find that the minimum distance between these sequences is three, and that, unlike some other such distances, it can be achieved in only one way. In many other cases, there are several possible paths. When this whole procedure is applied to all 351 pairs, there results a distance matrix analogous to the dissimilarity matrices employed in multidimensional scaling.[10]

10 To simplify exposition, we have omitted a final data manipulation. Since we do not want disparity in length to influence unduly the distances between sequences, these distances have all been standardized with respect to length. For each pair of sequences, the minimum cost transformation value is divided by the length of the longer sequence to remove length effects. The true distance in Table 3 is thus 1.0 (for the deletion of P1) + 0.8 (for the P8,P3 substitution) + 0.4 (for the R1,R2 substitution), all divided by 6, the length of the longer sequence; the result is 0.367.

ANALYSIS OF DISTANCES DERIVED BY OPTIMAL MATCHING

*Expectations* The distance matrix derived from this procedure may be analyzed by any standard method for distance data. Given our empirical interest in clustering within a cohort, it is particularly important to compare within-cohort distances to between-cohort distances. The various hypotheses about village traditions can be translated into expectations about these different distances. The strict tradition hypothesis, as expounded by Sharp, holds that there should be no variation either by cohort or by dance within cohort. There is a single village tradition, which each cohort should express. Since, however, there can be more or less attention to detail in a tradition, a strict tradition theorist might expect that some cohorts would display more internal variation than others. At the same time, however, the core of the tradition should always be the same so that, on average, mean distance within cohorts should not be substantially different from mean distance between cohorts. Both should be minimal but, in particular, distance between cohorts should be small. A theorist interested in cultural drift might think differently. At any given time, a tradition may be a unified phenomenon, but over time it may drift in various directions. Such a theorist would expect the difference between cohorts to be greater than the difference within them and would expect differences within to be fairly uniform from cohort to cohort.[11]

A revisionist theory would combine a belief in drift with a suspicion of the whole idea of village tradition. For the revisionist, the only native cohort in the Ilmington data was the 1887 to 1897 cohort. These were local dancers uninterested in the larger revival, reporting the dances as they had learned them from their predecessors. The 1887 to 1897 cohort is thus a benchmark cohort, in which the processes of tradition maintenance functioned independently of the morris revival. A revisionist would believe that reconstructionists artificially limited the variation of the dances, because of their *belief* in the unity of village traditions. On this argument, the distances within cohorts should be *inversely* proportional to the degree of revival involved. They should be *least* in the completely artifical 1867 data, and *most* in the completely unreconstructed 1887 to 1897 data. Bennett's revival of 1906 was

11 Forrest, *Morris and Matachin; idem*, "Here We Come A-Fossiling."

basically a traditional revival, since Bennett was a local dancer and musician reviving the dance after a hiatus, much as Michael Johnson and others had in 1887. But since Bennett was a single individual, who saw the tradition in his own particular way, a revisionist theory would expect him to exercise some narrowing influence. Internal variation in the 1906 to 1910 cohort should lie between the 1867 and 1887 extremes.

Variation among Bennett's later dances is more difficult to predict. The dances were directly based on Sharp's published work, yet had had time to drift by 1945, even though the drift may have been circumscribed by the published figure sequences. Thus, although the revisionist view would again expect internal variation in 1945 to lie between the figures for 1867 and 1887, it could not really predict whether the 1945 or 1906 cohort would show greater internal variation. It would, however, definitely expect within cohort distances to be significantly *smaller* than between cohort distances, since it would believe that drift occurs, and expect that three of the four cohorts would be significantly narrowed by revival influence.

*Within and Between Group Comparisons* In each matrix shown in Table 4, the figures are mean distances. The main diagonal comprises the mean distances within groups. The off-diagonal elements are distances between groups. Our original model was that 1867 and 1945 were "constructed" dances whereas 1887 and 1906 were "unconstructed." This model is shown in the first matrix. The mean distance within group is considerably less among the constructed dances than among the observed dances; the constructionists make traditions look more coherent than in fact they are. Furthermore, the distance between the constructed dances and the observed ones is substantial, more than twice the average within the constructed group, which implies that the constructionists assembled something different from the observed dances. Since the sampling distributions are obscure, we have tested the significance of the ratio of between- to within-group distance with a jackknife ratio test, and find t = 2.74, with 26 degrees of freedom. This indicates a significant difference, verifying the interpretation that the constructionists devised something different from the tradition that might have existed.[12]

12 The Jackknife procedure is a non-parametric approach to variance estimation em-

*Table 4* Distance Tables

| | 1867 AND 1945 | 1887 AND 1906 |
|---|---|---|
| 1867 and 1945 | .264 | |
| 1887 and 1906 | .637 | .492 |

t(26) = 2.74

| | 1867 AND 1945 | 1887 | 1906 |
|---|---|---|---|
| 1867 and 1945 | .264 | | |
| 1887 | .564 | .587 | |
| 1906 | .676 | .729 | .262 |

t(26) = 5.69

| | 1867 | 1887 | 1906 | 1945 |
|---|---|---|---|---|
| 1867 | .070 | | | |
| 1887 | .543 | .587 | | |
| 1906 | .686 | .729 | .262 | |
| 1945 | .452 | .607 | .656 | .380 |

t(26) = 5.63

| | SHEPHERD AND BUMPUS | BLACK JOKE | MAID OF THE MILL |
|---|---|---|---|
| SHEP-BUMP | .436 | | |
| BLACK JOKE | .549 | .512 | |
| MAID | .568 | .618 | .644 |

t(26) = 1.95

The second matrix in the table splits the two waves of unconstructed data and shows something startling. The within-group distances remain large in the 1887 wave. By 1906, the distances are much smaller and are on the level of the pooled constructed data. Once again, however, the distances between groups are substantial, and, again, a jackknife test supports the significance of the between-group differences, t = 5.69, with 26 degrees of freedom. The data seem to support the revisionist

ployed when underlying distributions are unknown and possibly not normal. Bradley and Bradley, "Application of Sequence Comparison," discuss the procedure applied here. A useful general reference is Bradley Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (Philadelphia, 1982). The method was originally proposed by Frederick Mosteller and John W. Tukey, *Data Analysis and Regression* (Reading, Mass., 1977).

theory that even a traditional revivalist like Bennett could limit the dances' range of variation considerably. Indeed, when we split out the 1945 cohort by itself (in the third matrix), we find that the purity that Bennett invented in response to Sharp apparently dissipated, for the distance within the 1945 cohort rises to .380.

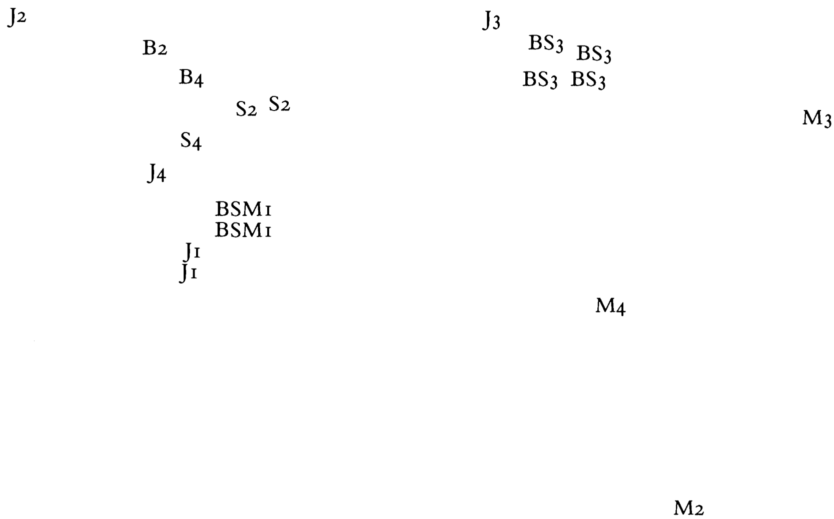
It is clear that reconstruction imposes a narrowing discipline on tradition, but that this discipline does not endure. Furthermore, the between cohort distances show that the dances drifted a good deal over the eighty years reported here. On the one hand, variation was at its greatest in precisely that data wave least touched by the revival. On the other, both its traditional successor (1906 to 1910) and the traditions invented to purify it were somewhat at variance with it, and with each other. The notion of a constant, steady tradition reaching into a remote antiquity seems, in short, a folklorists' dream. As Sam Bennett's two revivals illustrate, tradition itself is most likely a matter of continuous revival and drift, revival and drift.

*Analysis by Scaling* There are techniques available that allow us to envisage the mechanism that generates this drift. We have analyzed the distance matrix by using multidimensional scaling. This procedure produces the best possible n-dimensional representation of the data, along with a "stress" figure specifying how good "best" actually is. Figure 1 gives the two-dimensional scaling of this distance data, which has the acceptable stress value of .083. The figure provides a good, but not exact, representation of the information in the distance matrix. Each dance is identified by a letter and a number. The letter tells the dance, the number the cohort of the data. Multiple letters indicate identical dances.[13]

It is at once clear that the first constructed cohort (1867) is restricted, compared to all of the others. The second constructed cohort (1945) apparently owes its higher internal distance less to drift than to one dance, "Maid of the Mill." Bennett's 1906 to 1910 cohort spreads across about the same space, again extended by the wayward "Maid of the Mill." The 1887 traditional data are scattered across the page, although, again, deleting "Maid of

13 The classic references on multidimensional scaling are Roger N. Shepard, "Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function," *Psychometrika,* XXVII (1962), 125–140, 219–246; Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Non-metric Hypothesis," *ibid.,* (1964), 1–27. A useful recent review is Carroll and Arabie, "Multidimensional Scaling."

*Fig. 1* Two Dimensional Scaling of Figure Sequences




| | | | | | |
|---|---|---|---|---|---|
| B | = | Bumpus o'Stretton | 1 | = | 1867 |
| S | = | Shepherd's Hey | 2 | = | 1887–1897 |
| J | = | Black Joke | 3 | = | 1906–1910 |
| M | = | Maid of the Mill | 4 | = | 1945 |

the Mill" would leave a much smaller cluster. The surprising fact is that to some extent the clustering of the sequences is by *dance*. With the exception of the artifically homogenized first wave, the various versions of "Maid of the Mill" lie scattered around the lower right quadrant. The "Black Jokes," with the glaring exception of the 1906 dance, lie along a diagonal on the left-hand side. "Bumpus o'Stretton" and "Shepherd's Hey" occupy a block in the middle. The left/right dimension of this space is partly a dimension of length. But, other than that, the differences reflect real figure differences, and the grouping is at least partly a grouping by dances.

There is therefore historical coherence here, but not the historical coherence that the revivalists expected. The dances seem to shape even the supposedly common figures, not merely the distinctive chorus figures that distinguish them from one another.

This indicates that the local village traditions supposedly evidenced by the common figures were in fact affected by the particular *dance* traditions, which are shared by many villages. A jackknife test of the dance hypothesis, however, shows that this interpretation may not be correct. The figures are shown in the last panel of Table 4, and indicate a marginally significant difference.[14]

The summary implications of the various analyses are clear. The conception of an invariant tradition within the village community is theoretically implausible from the outset, and is, in fact, empirically rejected. The pattern found here is of a loosely bound tradition in which there is some steadiness, but no constancy. To be sure, the Ilmington data are merely one sample of the larger morris community, and we are continuing our investigation of other such data. But this first analysis indicates that there are no deeply constant village dance traditions that impugn the customary picture of a rapidly changing community in the nineteenth-century English countryside.[15]

STRENGTHS AND WEAKNESSES The methods used in this article are now widely available. They are an important supplement to methods currently available for sequence analysis. They apply both to recurrent event sequences and to non-recurrent (permutation) sequences, unlike seriation methods, which apply to the latter alone. Since optimal matching methods give evidence directly on sequences, rather than indirectly through evidence on events, they allow the *variation* among sequences to be analyzed. With seriation methods, since the underlying common sequence is found by scaling distances between events, there is no further

14 Concern about the stability of the results led us to several perturbation analyses. We ran the analyses with different indel weights, and also without "Bumpus o'Stretton," which seemed identical with "Shepherd's Hey." There were no changes in the pattern of group effects or in the scaling. The results thus seem stable, and their interpretability was established.

15 The real problem of representativeness concerns not spatial but temporal location. The data demonstrate flux in the dances, and over a substantial period, but that period comes after the heyday of morris dancing. There is little reason, however, to suppose that the mechanisms of transmission had substantially changed. Indeed, if there was a change, it was toward the stricter, more enforceable mechanisms of the revival. The observance of drift under revivalist conditions—written records and conscious control—guarantees it under the looser conditions characteristic of earlier times. It is therefore a fair conclusion that there are no constant traditions that impugn the standard picture of rapid rural change.

analysis possible. With optimal matching methods, since the distance matrix describes the sequences themselves rather than the events, a rich variety of clustering and scaling methods are applicable. One can discover more than one typical pattern, for example, or find branches in careers that start out similarly; one can scale the data and inspect them visually; and one can test hypotheses by employing cluster methods.

However, optimal matching methods have certain disadvantages. Since they employ only rank information about events, they generally reject continuous data even when these data are available. Seriation methods, by contrast, would retain this information about exact timing, although they would lose the scaling possibilities. They may thus be preferable for some data. However, it is possible to retain continuous information under optimal matching techniques by embedding sequential data in real time. One locates the event on a time line, then considers compression and expansion (rather than insertion and deletion) as elementary operations. (Substitution remains the same.) This continuous optimal matching is called time warping and can possess the advantages of seriation without its disadvantages.[16]

For historians and social scientists a more problematic weakness of optimal matching methods may be their basis on dyadic data. Most social science data are monadic—each case is defined by a single value on a given variable. But distance data such as those generated by optimal matching techniques are dyadic—data values are defined for pairs of cases. In the dyadic situation, the actual amount of data to be analyzed rises with the square of the number of underlying cases. Methods for analyzing such data must therefore restrict the number of cases studied. Familiar techniques for monadic data may easily study 1,000 cases or more. Dyadic data sets must be smaller. This weakness may, of course, prove a strength: where computational necessity requires restriction, users may choose their cases more carefully. The theoretical work necessary to make such choices will inevitably sharpen analysis.[17]

16 See Sankoff and Kruskal (eds.), *Time Warps*, "Pt. II: Time Warping, Continuous Functions, and Speech Processing," 121–209.

17 We are adopting the phraseology of "monadic" versus "dyadic" variables from Kruskal, "Overview;" in Sankoff and Kruskal, *Time Warps*, 36–37.

FURTHER EXAMPLES The strengths of optimal matching methods are best illustrated by some short examples of how they might be applied to other historical problems. There are a wide variety of possible applications and the two that follow are merely illustrative. But they show the kinds of issues that can be addressed by these methods.

As in our study of morris dance, optimal matching can be used to analyze the history of cultural symbols that have sequential structure. In this case, the symbol involved has a much greater historical import than does the morris dance; it is the Christian eucharist, investigated by Gregory Dix in magisterial detail. Dix's problem was to establish the "family tree" of the eucharist, the pattern by which innovations diffused from one or another center to reshape the underlying ritual structure established in the pre-Nicene church.

The eucharist's family tree can be established by investigating the different sequences of elements in the mass. Thus, the newer rites (to c. 800 AD) vary greatly in the material surrounding the greeting of the synaxis. For example, the Armenian rite has censing, entrance chant (*Monogenes*), greeting, psalm of the day, hymn (*Trisagion*), and litany before the lections begin; the Byzantine rite has censing, litanies, entrance chant (*Monogenes*), greeting, and hymn (*Trisagion*); the Milanese rite has entrance chant (psalm), litany or hymn (*Gloria* with *Kyries*), greeting, and prayer. Optimal matching offers an effective means of deciding how these (and other) rites are clustered, and which is probably descended from which. Dix himself analyzed the data by aligning them in a vertical table that showed exactly the transformations, insertions, and deletions that had taken place. Formal methods would offer a test of the hypotheses that he so laboriously examined by hand: that liturgical changes tend to originate in the east and move west; that Alexandria, in particular, influenced Roman developments; and so on. Indeed, careful analysis of Dix's data, applying multidimensional scaling to alignment-based distance measures, might reveal liturgical connections that Dix missed.[18]

Any cultural ritual or performance with a sequential organization can be subjected to such analysis. One can as easily

18 Gregory Dix, *The Shape of the Liturgy* (London, 1945; 2d ed.). The characteristic entry rites of the later period are shown on the chart bound opposite p. 432. Throughout his analysis, Dix uses a phraseology that presages matching techniques.

analyze the evolution of the sonata form in music as that of morris dances or religious liturgies. In each case the data must first be placed into a standardized format of sequences of well-defined elements. The costs for replacement, insertion, and deletion of these elements must then be set. Using these costs, optimal matching methods may then be used to create distance matrices. These matrices can then be employed to find contemporary clusters or to test possible evolutionary trees relating the various sequences. The procedure is in every case the same as that followed in our detailed study above.

The methods are equally applicable to sequences of actions, such as the riots with which we began. The use of optimal matching techniques to create such typologies of action sequences is their most general one. An example is a study applying these methods to histories of the adoption of welfare programs by various nation-states.

In the standard view, there are five basic welfare programs—workmen's compensation, sickness/maternity benefits, old age/incapacitation/death supports, family allowances, and unemployment insurance. In what order are these programs adopted, and why? Cutright inspected data on seventy-six countries and found that sixty-three of them had concentric samples of the list just given: that is, their policies included workmen's compensation and up to four more policies *without skipping over any program in this ordered list*. This Guttman scale quality led him to infer that the policies were normally adopted in the order just given. Yet inspection of the actual dates of adoption shows that this inference was not correct.[19]

Optimal matching offers an effective way both to find the true typology of adoption sequences and to uncover the reasons why different countries fall where they do in that typology. Since we do not wish to lose the continuous time information present in the dates of adoption, we must first embed the data in continuous time. This step is easily accomplished. There are thirty-two possible situations for a country to be in, since there are five policies and a given country may have or lack any of them. We find which of these situations each country has maintained during

19 Phillips Cutright, "Political Structure, Economic Development, and National Social Security Programs," *American Journal of Sociology*, LXX (1965), 537–550. The study here discussed is being conducted by Stanley DeViney and Abbott.

each ten-year interval since its first adoption of a welfare program (or since some arbitrary prior point). The sequence of these situations, for each country, describes the historical sequence of its general welfare system. For example, France acquired its first old age benefits in 1910, its first maternity/sickness program in 1928, workmen's compensation in 1898, unemployment insurance in 1905, and family allowances in 1932. It thus maintained the situation "compensation alone" for most of the first decade after 1898; then the situation of "compensation + unemployment + old age" for most of the second and all of the third decades (1908 to 1928); and then the situation of "all programs" for most of the fourth and all the subsequent decades to the present. Unlike the sequences implicit in data on riots, such sequences exist in real time rather than in rank ordered time. If a country stays in one place for twenty years, that situation contributes two elements to the sequence, not just one.[20]

The sequence data so generated can be analyzed with optimal matching. As in any such application, the first task is to create insertion, deletion, and substitution costs. Given the transformation to combination situations, it seems natural to define substitution costs in terms of the number of common elements that two situations share. The combination of "workmen's compensation + sickness/maternity benefits" shares one of its elements with the situation of "workmen's compensation," and so we might say that the two situations, sharing one out of a total (in both) of two properties, are .50 alike. A situation lacking only one property would be .80 like the situation possessing all properties. One could subtract these likenesses from 1.0 to get a substitution cost. Insertion costs are rather more difficult to specify, but can be set by a similar process. Given these costs of insertion, deletion, and substitution, we can then proceed to the second and third steps of analysis in order to create a typology of welfare histories. We first apply optimal matching to create the

20 Implicit in this procedure are numerous assumptions about the nature of the data involved. What does first adoption really mean: The first comprehensive law? The first partial law? The first law in a particular state (in federal systems)? The date that comprehensive coverage became "a reality"? Such questions are analyzed in detail by Abbott, "Event Sequence and Event Duration: Colligation and Measurement," *Historical Methods,* XVII (1984), 192–204. The information on France comes from *Social Security Programs throughout the World, 1981* (Washington, D.C., 1982).

distance matrices, and then apply cluster analysis or multidimensional scaling to find the basic shape of the typology.

The final step in such an analysis is to discover the variables shaping that typology. Our exemplary case above did not include such an analysis; we interpreted the scaling results by inspection. But there is a formal method. It is customary in scaling and clustering to test the variables that are thought to generate the observed pattern by regressing them on the scaling coordinates or cluster memberships. In this case, the variables to be considered are those specified by the major theories of welfare development. There are three such theories: one attributes the evolution of welfare programs to general processes of development; another, to the rise of the working class; a third, to the structure of the state. In the first case, the important variables are economic development and the age of the population; in the second, the dominance of working-class parties and the proportion of labor force unionized: in the third, the power of the state and its degree of centralization. To find which of these variables underlies the observed scaling of adoption sequences, we treat these variables as *dependent* ones, and try to predict them (one by one) with the scaling coordinates. If we find one or two particularly strong relationships—with multiple correlations in the .80 or .90 range—we can assume that the variable so predicted was an important underlying dimension of the adoption sequence space, and hence an important determinant of adoption sequences themselves.[21]

21 This use of regression for interpreting the dimensions of a scaling result is standard in the literature. For a general discussion, see Kruskal and Myron Wish, *Multidimensional Scaling* (Beverly Hills, Ca., 1978), 35ff. The most commonly used software for the procedure is PROFIT. See J. J. Chang and Carroll, "How to Use PROFIT, a Computer Program for Property Fitting by Optimizing Nonlinear or Linear Correlation" (Bell Laboratories, Murray Hill, N.J., 1968). A recent published example of this technique is Susan C. Weller, "New Data on Intracultural Variability: The Hot-Cold Concept of Medicine and Illness," *Human Organization,* XLII (1983), 249–257.

On the modernization theory see, e.g., Cutright, "Political Structure"; Robert W. Jackman, *Political and Social Equality* (New York, 1975); Harold Wilensky, *The Welfare State and Equality* (Berkeley, 1975). On the working-class strength thesis, see Frances G. Castles and R. D. McKinlay, "Public Welfare Provision and the Sheer Futility of the Scoiological Approach to Politics," *British Journal of Political Science,* IX (1979), 157–172; John Myles "Comparative Public Policies for the Elderly," in Anne-Marie Guillemard (ed.), *Old Age and the Welfare State* (Beverly Hills, Ca., 1983), 19–44; Michael Shalev, "The Social-Democratic Model and Beyond," in Richard F. Tomasson (ed.), *Comparative Social Research* (Greenwich, 1983) VI, 315–352. On the state centralization thesis, see Margaret Weir and Skocpol, "State Structures and Social Keynesianism," *International*

Such an illustration shows a more general use of optimal matching techniques. They can not only classify sequential data, but can also investigate the causes determining that classification. In more traditional historical language, not only can they find generalized narratives; they can also locate the motive forces behind those narratives. The application to welfare adoption sequences shows how potentially powerful this technique can be. In short, optimal matching is an important new technique for analyzing sequence data. It is applicable in a wide variety of historical settings, but particularly where cases consist of careers of (possibly) repeated events, drawn from a relatively small universe of events whose substitutability can be specified. Studies of revolutions, of negotiations, of international politics, or of development can all benefit from the application of such techniques, as can studies of the history of sequential cultural data.

---

*Journal of Comparative Sociology,* XXIV (1983), 4–29; DeViney, "Characteristics of the State and the Expansion of Public Social Expenditures," in Tomasson, *Comparative Social Research,* VI, 151–173.