

Is Optimal Matching Sub-Optimal?

Matissa Hollister
Department of Sociology
6104 Silsby Hall
Dartmouth College
Hanover, NH 03755

matissa.n.hollister@dartmouth.edu
office: (603)646-3524
fax: (603) 646-1228

Keywords: Optimal matching, sequence analysis, careers

Is Optimal Matching Sub-Optimal?

Abstract

Optimal matching (OM) approaches sequences in a way that is intuitively and theoretically appealing. Two problems exist, however, with the ways in which OM has been used to date. One problem stems from the method itself. I point to a flaw in OM “indel costs” and propose a solution to this flaw. The second problem with OM is the need for benchmarks to examine the added value of OM techniques and to determine whether one version is better than another. To that end, I provide an empirical test of traditional OM, an alternative “Localized OM,” and a third, more simplified method (“Sequence Comparison”). I document the identified problem with traditional OM and show that this problem is solved by Localized OM. However, neither version of OM fares substantially better than the more basic Sequence Comparison. These results raise questions about the method as a whole and point to the need for further assessment and the development of appropriate benchmarks.

In the late 1980s, Andrew Abbott introduced a technique known as Optimal Matching to the social sciences. Optimal Matching (OM) is a method borrowed from biology that can be used to create a measure of similarity (or more often dissimilarity) between pairs of sequences, for example two individuals' work histories. OM results are frequently used to identify clusters of similar sequences to develop, for instance, a typology of careers. Appendix D summarizes recent publications using OM. In the past two years researchers have also developed OM modules for several of the mainstream statistical software packages (e.g. Brzinksy-Fay, Kohler, and Luniak 2006), which will likely increase OM use in the future.

Optimal matching is just one of several types of sequence analysis algorithms (methods for comparing pairs of sequences). Abbott adopted OM directly from biology, purposely choosing a simple algorithm to make the method "attractive and accessible" (Abbott 1995: 233).¹ Abbott invited researchers to explore alternate configurations of the algorithm, but few people have followed this suggestion; almost all applications of sequence analysis in the social sciences have used the OM algorithm. These studies have led to several interesting and valuable insights, but one is left wondering whether this traditional OM method provides the best results. Are there better ways to configure the OM algorithm? Does OM provide better results than other methods? What do we mean by "better"? These are the questions I explore in this paper.

I focus in particular on two problems with how OM has been used to date in the social sciences. The first problem stems from the OM algorithm itself, in particular the way it uses "insertions" and "deletions." I propose an alteration to the OM algorithm, which I call Localized OM, that takes an approach that I argue is more appropriate for the analysis of occupational careers. The second problem with OM is the lack of clear benchmarks that can be used to test the results. I discuss several possible benchmarks and suggest a new possibility: OM results should be correlated with future outcomes. In the final section of the paper, I present results from an empirical example that uses this benchmark to test

traditional OM, Localized OM and a third approach, Sequence Comparison, which is less computationally intensive than the two OM algorithms.

The results from the empirical example point to several interesting findings. First, I find that there is indeed a problem with insertions and deletions using the traditional OM algorithm and I find that Localized OM is successful in addressing this problem. I also find, however, that both traditional and Localized OM seem to provide only marginal improvements compared to Sequence Comparison. These findings suggest the need for further scrutiny and testing of OM as a social science method.

Measuring sequence dissimilarity

The basic goal of OM is to measure the level of dissimilarity between a given pair of sequences. Consider two sequences, A and B . Each sequence is composed of a series of elements measured at each time period.

$$\begin{aligned} A &= (a_1, a_2, a_t \dots a_p) \\ B &= (b_1, b_2, b_t \dots b_p) \end{aligned} \tag{1}$$

where:

a_t = sequence element at time t for person A

b_t = sequence element at time t for person B

p = number of time periods²

Each element in a sequence is comprised of a certain “state” selected from a set of possible states:

$$\begin{aligned} a_t &\in S = \{s_1, s_2, s_3 \dots s_m\} \\ b_t &\in S \end{aligned} \tag{2}$$

where:

m = number of possible states

The set of states used to represent the sequences is determined by the researcher and obviously has an impact upon the eventual results. For instance, the states might be Erikson-Goldthorpe class

categories (Halpin and Chan 1998), the number of jobs held in a given year, one's residential location (urban/rural/suburban) (Stovel and Bolan 2004), marital status, or even a combination of these different factors. As with other forms of data analysis, the selection of states in OM reflects the interests and hypotheses of the researcher.

Once we have decided how to express the two sequences A and B as a series of specific states, how do we decide whether the two sequences are similar or different? The strictest criteria would be to consider the two sequences to be similar only if they are identical. For instance, one might identify the specific sequences that occur the most frequently in a dataset, which in effect only considers sequences to be similar if they are identical. This approach works well if the number of possible permutations of sequences are small (both a small p and a small m).

Rather than requiring identical sequences, a second possibility is to compare a pair of sequences and count the number of time periods where the elements match as a proportion of the total number of elements. Pairs of sequences with a high proportion of matching elements would be considered more similar. This is the idea behind measures such as the Jaccard Coefficient and Hamming Distance.³ This approach works better than the previous one for long sequences (larger p 's), but the method still requires a small set of possible states (small m). With a large set of states the number of matched elements would be very low.

In the following sections I consider two ways to add considerably more complexity and flexibility beyond the basic methods described above. The first addition is the ability to examine sequences comprised of much larger and complex sets of possible states ("variable substitution costs"). Incorporating this first addition leads to a method I will call *Sequence Comparison. Optimal Matching*, meanwhile, incorporates both variable substitution costs and a second addition: the ability to shift the sequence elements in time for better comparisons.

Variable substitution costs

The methods discussed above require sequences composed of a limited set of possible states (small m). In cases where m is large, one solution is to designate some states as closer to each other than others. When comparing two states, therefore, they do not necessarily have to be an exact match, they could just be similar states. In the case of sequences composed of occupations, for instance, we may want to designate carpenters and plumbers as more similar to each other than carpenters and lawyers. In the language of OM, this measure of similarity or difference between states is called *substitution costs*. The methods described above use what might be called “unit substitution costs”: all states are equally distant from all other states (either two states are the same or they are not). Instead we can designate “variable substitution costs.”

Variable substitution costs involves specifying the distance between each pair of possible states in S . This creates a symmetric $m \times m$ matrix of substitutions costs:

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mm} \end{bmatrix} \quad (3)$$

where:

w_{ij} = substitution cost between two states: s_i and s_j

$w_{ij} = w_{ji}$ (the substitution matrix, W , is symmetric)

When comparing two sequences, we can add up the substitution costs between the elements at each time period to create a measure of dissimilarity. I will refer to this method as *Sequence Comparison*.

$$D_{SC} = \sum_{t=1}^p w_{a_t b_t} \quad (4)$$

where:

D_{SC} = Sequence Comparison dissimilarity measure

$w_{a_t b_t}$ = substitution cost between the states in sequences A
and B at time t

The additional flexibility of variable substitution costs also brings with it some challenges, namely the need to specify W , the substitution costs between every pair of states in S . Several approaches have been used to set these costs:

Theoretically derived. Early OM applications commonly relied on theory and the researcher's judgment to set substitution costs. For instance, Halpin and Chan (1998) assigned occupations to the Erikson-Golthorpe class schema and used the hierarchy of class groupings in that schema to assign costs. Theoretically derived substitution costs have run into criticism, however. Theory rarely provides a precise indication of how these costs should be set, and so the process is often viewed as subjective and arbitrary.

Scalable values. A second possibility is to assign quantitative values to each state. For instance, if the states are occupations one could use Duncan SEI scores (Duncan 1961). The substitution cost between two occupations would then be the absolute value of the difference in their SEI scores. The problem with this approach is that it requires the values to be aligned on a one-dimensional scale. So, for instance, a secretary and an electrician have the same SEI score and therefore would be considered identical. One of the advantages of using variable substitution costs is that it does not require the use of values on a one-dimensional scale. The matrix of substitution costs W allows the researcher to specify a specific distance between each pair of states. Therefore, distances between states do not have to be additive. In some cases scalable values, or modifications of this approach, are appropriate, particularly if the states are intrinsically quantitative. For instance, Stovel (2001) examined patterns in incidents of lynchings. Since her sequences states were counts of the number of lynchings in a given time period, her substitution costs were calculated directly from these counts.⁴

Transition rates. The most common technique in more recent OM studies (see appendix D) is to use transition rates. This approach uses the data to determine the similarity between two values, thus avoiding accusations of subjectivity. Transition rates measure the proportion of people in state s_i in time t who are in state s_j in time $t + 1$, as well as the proportion of people in state s_j in time t who are in state s_i in time $t + 1$. Two states are considered similar if transitions between them are common.⁵

The main limitation of transition rates is that it requires either a small set of possible values or a large dataset to measure transition rates. For example, sequences consisting of 300 detailed census occupation codes would involve measuring the transition rates between 44,850 different occupation pairs. One would therefore need a very large dataset to get an accurate and fully-detailed measure of transition rates.⁶ Due to these issues, most studies using OM have limited themselves to a very small set of possible values by either limiting the setting (e.g. workers at a bank) or by simplifying the data (e.g. coding occupations into broad class categories).

Composite states. A number of researchers have examined sequences composed of what I call composite states. For instance, Stovel, Savage, and Bearman (1996) examined career sequences at Lloyds Bank. They identified two dimensions behind the positions at the bank: job title and branch location. They created five categories of job titles and six categories of branch locations and used transition rates to create substitution costs between the different job titles and a separate set of substitution costs between the branch location positions. A given state in a sequence was a combination of both title and location. The substitution cost between two states would be the sum of the title substitution cost and the location substitution cost. A number of other researchers have examined similar sequences composed of composite states with separate substitution costs for each dimension of the state and the final substitution costs equal to the sum of these separate substitution costs (Aassve, Billari, and Piccarreta 2007; Pollock 2007; Williams and Han 2003). The substitution costs for these

underlying dimensions were set using any of, and often using combinations of, the above methods (theoretically derived, scalable values, and transitions rates).

The complex state approach is in effect a way to greatly increase m , the number of possible states, since each state can be any combination of the underlying dimensions. Creating separate substitution costs for each dimension and adding them together is a simplified alternative to specifying a very large substitution cost matrix W with m equal to all possible combinations of the different states.

A DOT-based measure for occupations. The empirical example at the end of this paper demonstrates an approach that creates a substitution cost matrix between detailed occupation codes. The Dictionary of Occupational Titles (DOT) measures a wide variety of occupational characteristics including training time, physical demands, job characteristics, and temperaments (National Academy of Sciences 2001). One can compare a pair of occupations across all of these measures and create a summary value for distance between each pair. In effect this measure is a composite approach, using quantitative measures of multiple dimensions of occupational characteristics to create a more dynamic and detailed measure of differences between pairs of occupations than previous options such as Duncan SEI scores or Erikson-Goldthorpe categories. An additional advantage of the DOT-based substitution costs is that they are based upon externally created data and therefore, unlike traditional transition rate measures, they are not dependent upon the specific sample being used.

Shifting time: indel costs

The addition of variable substitution costs, as just described, adds a certain amount of complexity in comparing sequences. Once the substitution costs are set, though, the actual calculation of differences between a pair of sequences is relatively straightforward. As shown in equation 4 above, the Sequence Comparison dissimilarity measure, D_{SC} , can be calculated by summing the substitution costs between elements at each time period.

The Optimal Matching algorithm adds another layer of complexity. It uses variable substitution costs but then asks an additional question: would these sequences look even more similar if we could shift all or part of a sequence in time? The OM algorithm considers whether inserting an extra element, or deleting an element, in one or more locations in a sequence would line up the remaining elements better and reduce the substitution costs. Inserting an element into one sequence (shifting all or a portion of the sequence to the right) is basically the same as deleting an element from the other sequence (shifting it to the left). So, in OM terms insertions and deletions are collectively referred to as *indels*. Similar to substitution costs, shifting sequence elements in time should have a “cost” c_i . The dissimilarity measure for Optimal Matching, therefore, is the sum of the indel costs and the substitution costs in the newly aligned pair of sequences:

$$D_{OM} = \min \left(\sum_{i=1}^k c_i + \sum_{t'=1}^{p'} w_{a_t, b_{t'}} \right) \quad (5)$$

where:

c_i = cost of shift (“indel” cost)

k = number of shifts

t' = time in newly aligned sequences

p' = number of elements compared in new alignment

$w_{a_t, b_{t'}}$ = substitution cost between elements at t'

The challenge with this OM dissimilarity measure is that there are many different possible alignments. The OM algorithm, therefore, considers all possible alignments and identifies the alignment that minimizes equation 5 above. Basically, it considers the trade-off of the cost of a given indel and the savings this shift might have for the substitution costs. There may in fact be more than one “optimal” alignment. The actual specific alignment is not important, the main output of OM is the minimized D_{OM} .

value. The process of finding this minimized value is an iterative process that requires significant computing resources (although the feasibility of OM with large numbers of sequence pairs has improved significantly over the last two decades with advances in computer technology). See Abbott and Hycak (1990) and Sankoff and Kruskal (1999) for more details on the mechanics of the OM algorithm.

Sequences with unequal lengths

So far the description of OM has used the example of sequences of equal length (both length p). One advantage of OM over Sequence Comparison and other methods is that it offers the opportunity to compare sequences of unequal lengths. The OM algorithm uses indels at optimal locations to “stretch” the shorter sequence to the same length as the longer sequence. Sequences of unequal lengths, though, pose some challenges to setting OM costs, particularly indel costs, that have not been completely resolved. See appendix A for a brief discussion of these issues.

Alternate approaches

Dijkstra and Taris (1995) have proposed an alternate method for measuring differences between sequences and Elzinga (2003) expanded on their algorithm. Unlike OM, these methods are able to recognize situations where two sequences contain similar values but these values do not align because they are either extremely shifted in time or because they are in a different order.⁷ The authors showed evidence that these methods provide better results than unit-cost OM (OM that uses indels but does not have variable substitution costs).

A central problem with the Dijkstra & Taris and Elzinga algorithms, though, is that they do not allow for variable substitution costs. Given the widespread use of variable substitution costs in sociological applications, this is a serious limitation. Nevertheless, researchers examining sequences composed of a small set of states might find that the gains from these alternate methods offset any losses from imposing a uniform substitution cost.

The Sociological Perspective

Now that we have a sense of how OM works, we might ask whether it is a useful method for social science research. Later in this paper I will examine empirical tests of the method. For the moment I focus on theoretical justifications for OM with a particular focus on the study of careers.

Proponents of OM usually point to two rationales for why OM is a valuable tool for the study of careers. The first argument is that OM allows researchers to study careers as an entire sequence. This approach is usually contrasted with event history analysis, which models careers as evolving one job at a time, the outcome of a series of transitions. There are a number of specific situations where this whole-sequence view would likely be appropriate, for instance job ladders within internal labor markets or preset training sequences within professional and craft occupations. One could also argue that the whole-sequence approach to careers is applicable more broadly because workers tend to follow understood “scripts”:

There are reasons to believe that individuals themselves try to structure their work histories into careers that they find culturally acceptable, into patterns they recognize (Abbott and Hycak 1990: 147).

OM approaches careers with this perspective, looking at job sequences as coherent wholes rather than a series of stochastic events.

A second argument for OM is that it is simply a good pattern identification algorithm (Abbott 2000). Even if careers are primarily generated by event history-like transitions, certain paths through these transitions may be more common than others and OM may be able to identify these common paths. One could then use other methods such as event history analysis to examine the processes that create these patterns.

Skeptics of the method question whether OM really is a generic pattern identification method. A number of people believe that the success of OM in biology, where it was originally developed, is not

due to its pattern recognition capabilities but rather due to the fact that the OM operations (substitutions and indels) mimic evolutionary processes. They then question whether these operations have a similar parallel in social contexts:

My skepticism stems, in part, from my inability to see how the operations defining distances between trajectories ([substitutions] and indels) correspond, even roughly, to something recognizably social (Wu 2000: 46).

Abbott (2000) argues against this position by pointing out that the parallels in biology are not as direct as they might appear and, therefore, that OM is indeed just a general search program.

I believe, however, that there are reasonable and useful interpretations of the OM operations for the social context. The goal of OM is to measure the extent to which two individuals share a similar underlying career pattern while allowing for some variation. Substitutions and indels allow for two common types of variation. First, individuals might hold slightly different jobs in a given year (substitutions). Second, individuals may move through a stage in a career path somewhat faster or slower than is typical, requiring the stretching or compressing of certain time periods using indels. In contrast to Abbott, therefore, I believe that the interpretations of these OM operations are relevant and useful. As I will show in the following section, these interpretations point to a potential problem in the OM algorithm when applied to careers.

Two issues with optimal matching

Thus far, this paper has focused on ideas and issues that have been broadly discussed in the existing literature. For the remainder of the paper, I focus on two issues that have gained less attention from previous researchers. The first issue involves a problem with the OM algorithm itself and the structure of indel costs. The second issue is the need for benchmarks to test OM results.

Insertion/deletion costs

Setting indel costs

Up to this point I have conveniently avoided the question of how one determines the cost for insertions and deletions, an omission that in part reflects the tenor of the existing literature. Articles often dedicate several paragraphs to the discussion of substitution costs and report the indel cost only as a footnote or in an appendix, if at all.

This lack of attention to indel costs is somewhat surprising because, as discussed above, insertions and deletions are a unique and central part of optimal matching. In addition, setting indel costs in the context of variable substitution costs can be tricky. The indel cost is traditionally a fixed value expressed as a proportion of the maximum substitution cost (the maximum value in the substitution cost matrix W). One important point to note is that any substitution cost can be eliminated with a combination of a deletion and an insertion (delete the unwanted element, insert the new element into the same spot). Initially, researchers sought to avoid these “pseudo-substitutions” by setting the indel cost to at least half the maximum substitution cost, thus ensuring that two indels would always cost more than any substitution. Early work in the natural sciences further suggested that the indel cost should be greater than the maximum substitution cost (Bradley and Bradley 1999), a convention that was adopted in early uses of OM in the social sciences (Abbott and Hycak 1990).

In more recent work, however, Abbott has suggested that indel costs should be set much lower. Abbott and Tsay (2000: 12) even stated that if indel costs “are set to any cost greater than half the largest substitution cost, indels will *never* be used” (a point repeated in Macindoe and Abbott (2004)). This statement, unfortunately, is not correct. High-cost indels will be used if they can save substitution costs across several elements with a better alignment.⁸ Regardless of the accuracy of this statement, though, Abbott raised an important question about the most appropriate indel cost setting. His

OPTIMIZE software program has a module that allows for visual inspection of the alignments of sample pairs of sequences. Abbott commented that:

Using this module, Abbott and colleagues ascertained that indel costs in the vicinity of 0.1 times the largest substitution cost tend to pick up the sequence regularities that appeared to be substantively interesting (Abbott and Tsay 2000: 13).

Few researchers seem to have followed Abbott's advice thus far. Most of the studies reviewed in appendix D have continued to set indel costs at values at least half of the maximum substitution cost. However, Abbott's call for extremely low indel costs has been repeated in the *Handbook of Data Analysis* (MacIndoe and Abbott 2004) as well as the documentation for an OM module in Stata (Brzinsky-Fay et al. 2006). It is likely, therefore, that future studies will follow this advice.

Setting indel costs at such a low level presents some problems, however. As mentioned above, one can always use a pair of indels instead of a substitution. With an indel cost set to 0.1 times the maximum substitution cost, OM will choose to use a pair of indels in the place of any substitution greater than 0.2 times the maximum substitution cost. In effect, all substitution costs greater than 0.2 times the maximum will be reset to this value. This process throws out much of the careful consideration a researcher puts into creating substitution costs in the first place. These considerations should be balanced against Abbott's conclusion that low indel costs achieve better alignments.

At the other end of the spectrum, Lesnard (2007: 11) has recently argued that indels "warp time" and should be "seldom and carefully used." His analysis, therefore prohibits the use of indels completely. Lesnard bases these assertions upon a theoretical belief that social phenomena are tied to their context in time, and that shifting these phenomena in time violates their fundamental character. It appears that there is broad disagreement, therefore, on setting indel levels.

The “free ride” problem

The few discussions to date on OM indel costs have focused on the question covered in the previous section: what fixed value should be used for the indel cost? Much less attention has been paid to a second problem, which relates to how insertions and deletions should operate within a social context.

Within a biological context, insertions and deletions do a fairly good job of imitating evolutionary processes. It appears that in biology, the fact that a strand of DNA has been cut is more important than the values of the elements that are inserted into that cut. Thus indels are said to add “padding” or “unimportant residues” (Lesnard 2007: 8) to the sequences.⁹

This view of indels is a crucial area where the sociological conception of careers diverges from biological approaches to sequences. As discussed earlier, indels in the career context might be seen as stretching or compressing periods of time within the sequence. Unlike in biology, we do care about the element that is inserted; inserting “accountant” between two periods as “farmer” should cost more than inserting “accountant” between two periods as “bookkeeper.” The traditional fixed-cost indel system, therefore, creates what I call the “free ride” problem: once the decision is made to insert or delete an element, the OM algorithm doesn’t care what element is inserted. It allows a highly unusual element to be inserted without any additional cost. A small number of researchers (Abbott 1995; Abbott and Hyrcak 1990; Lesnard 2007; Scherer 2001) have noted this problem but then quickly dismissed it as too computationally challenging.

Stovel (2001) offers one alternative approach to setting indel costs in her study of historical lynching patterns in the South. The sequences in the study were comprised of the counts of lynchings in each time period in an area. The indel cost was set to the value that was being inserted/deleted. Therefore, if a count of 5 was inserted into a sequence the indel cost would be 5. This approach in effect says that when we create a gap with a shift in time we should assume that the gap had zero lynchings

(regardless of the values of its neighbors) and if not we will charge more. This approach may make sense with certain types of sequences, but in the case of career sequences I suggest an alternate approach.

A localized indel approach

I propose an alteration to the OM algorithm that will address the two indel cost problems discussed above. This new algorithm will allow the researcher to set indels at lower levels, as suggested by Abbott, without the risk of a pseudo-substitution. In addition, the new algorithm reflects a more sociological treatment of insertions and deletions.

Consider the following pairs of sequences (with scalable values).

Sequence A: **1 1 99 1 1 10 40 90 1**

Sequence Z: **1 1 1 1 10 40 90 1 1**

Sequence B: **1 1 1 1 1 10 40 90 1**

Sequence Z: **1 1 1 1 10 40 90 1 1**

Under traditional OM, the dissimilarity between sequences A and Z would be the same as the dissimilarity between sequences B and Z; each pair can be aligned using one pair of indels. In fact, in the case of sequences A and Z, traditional OM would take advantage of the need for an insertion to slip in an extreme value (99) that would otherwise require a large substitution cost.

The new approach I propose would distinguish between these alignments. The new “localized” indel cost is composed of two parts. The first part is the same as traditional OM, it charges a fixed baseline cost for shifting the elements. The second part averages the distance between the newly inserted value and its two neighbors. The algorithm therefore charges a low cost for inserting or deleting values that are the same or similar to their neighbors (expansion or compression of time) while punishing attempts to use indels to insert values that are different from its neighbors. The equation for this new indel cost is:

$$C_{izj} = x \cdot w_{\max} + y \cdot \frac{(w_{zi} + w_{zj})}{2} \quad (6)$$

where:

C_{izj} = the cost of inserting state s_z between states s_i and s_j ¹⁰

w_{\max} = the maximum substitution cost

w_{zi} = the substitution cost between values s_z and s_i

w_{zj} = the substitution cost between values s_z and s_j

x = "time cost" (set by researcher)

y = "local cost" (set by researcher)

The first part of the equation is a fixed indel cost identical to traditional OM. It is set to some proportion, x , of the maximum substitution cost. The second part of the equation is the average of the distance between the inserted value and its neighbors multiplied by a factor y .

This new indel approach appears at first to complicate the issue. We now have two parameters, x and y , to set instead of just one. The interpretations of the parameters, though, are more apparent. The x parameter ("time cost") measures the extent to which one wants to preserve time. A low value allows the elements to be shifted in time with little cost, while a high value gives priority to preserving time. The y parameter ("local cost") indicates how much one wants to punish the insertion or deletion of unusual values, the importance of the local context. Traditional OM is a special case where $y = 0$, in other words local context does not matter and there is no punishment for inserting unusual values. In contrast, the combination of a low time cost (x) and a high local cost (y) results in a high cost for inserting or deleting unusual values, but a low cost for expanding or compressing time by inserting or deleting values that are very similar their neighbors.

The researcher is still left to decide how to set these two parameters, x and y . One of the goals of this new approach is to prevent OM from using a pair of indels in lieu of a substitution. The following formula, therefore, should serve as a guideline:

$$1 - 2x \leq y \quad (7)$$

As long as this equation holds, a pair of indels will never be used as a pseudo-substitution. See appendix B for the derivation of this equation. For further guidance in setting the indel parameters, we must turn to a second issue in optimal matching.

The benchmark question

Faced with several options for measuring dissimilarity between sequence pairs (Sequence Comparison, traditional OM, and Localized OM), as well as many options for setting different parameter levels, the question is how to determine whether one approach works better than another. In biology, researchers have the advantage of pairs of sequences where the correct alignments are known. Researchers can therefore test which parameter settings find the correct alignments (Vingron and Waterman 1994). In the social sciences we are not so lucky; there are no known correct answers. We therefore must turn to other types of benchmarks.

Levine (2000) criticized OM researchers for relying too much on surface validity, taking a “looks good, makes sense” (p. 35) approach. The value of OM is implied by the fact that the analysis provides interesting insights. There is often less effort, though, to demonstrate that the researcher has chosen the best parameter settings, or to demonstrate that OM provides better insights than alternate methods. Abbott’s advice for choosing indel parameter levels similarly relies on subjective evaluations of validity. “To determine appropriate levels, it is necessary to actually watch the effect of changing the indel level on a number of alignment pairs” (Abbott and Tsay 2000: 13). The “explore” mode of Abbott’s OPTIMIZE software program is designed to facilitate this process. This approach to setting indel levels is clearly subjective and a target for criticism. One could potentially improve on this approach by using

multiple coders to subjectively rate the similarity of pairs of sequences. These ratings could then be used to test the extent to which OM captures commonly-held conceptions of sequence similarity.

Another possible method for testing OM would be to conduct simulations. One could start with a set of sequences, introduce certain perturbations to those sequences, and then measure the extent to which OM is able to group the altered sequences back with their original “parents.” The challenge with this approach, though, is choosing appropriate perturbations. There will be an inevitable tendency for the researcher to select perturbations that implicitly favor his or her preferred technique. For instance, advocates of Elzinga’s (2003) method discussed earlier would likely include “swaps” (switches in the order of elements) as one of the perturbations since that method is designed to capture these swaps. On the other hand, one may believe, based upon theoretical or observational grounds, that swaps are unlikely to occur in real life and therefore should not be included in the simulation. One’s beliefs about careers, therefore, will affect both the selection of a method and the selection of simulation parameters, with the inevitable tendency to set up a simulation that favors the selected method. Without knowledge of how actual variation in careers occur, simulations are likely to be biased.¹¹

An additional possible benchmark is whether the OM results, once fed into a clustering program, identify homogenous clusters that are relatively easy to interpret. This benchmark is based upon a theoretical assumption behind many OM studies: that there are common patterns underlying the sequences that are waiting to be identified. If OM fails to find clear clusters this failure would indicate that either the theory is wrong (clear groups do not exist) or that the algorithm has failed to identify the patterns. Failure on this benchmark, therefore, may be due to problems with the theory rather than the method, but either way the failure might lead a researcher to conclude that the method is not very useful.

A final benchmark option is to measure the extent to which the level of dissimilarity between two sequences is correlated with differences in their states in the future. This benchmark is based upon

the assumption that career patterns exist and that it is difficult for an individual to switch career types. Again, this is a theoretical question that should be tested with the data. It is possible that the sequence analysis methods do a good job at identifying patterns, but that these patterns are not as important and path dependent as we think they are. Similar to the homogeneous clusters benchmark, therefore, failure on this benchmark may reflect either a failure of the method or a failure of the theoretical assumptions. I would argue, however, that this benchmark is a test of one of the more fundamental assumptions behind the use of sequence analysis methods: that these patterns will tend to persist into the future and will therefore have long-term impacts on individual outcomes. If the patterns are not persistent, if it is relatively common to switch from one to another or if they have little impact on future outcomes, then their interest as a subject of research is greatly reduced.

A valid criticism of this benchmark is that it is based upon the assumption that different patterns lead to different endpoints. It is quite possible that in some cases different patterns may lead to the same endpoint, something that would not be captured by this benchmark. I would argue, though, that in the case of occupational careers, especially with a diverse sample such as the one used in the empirical example below, one would expect the majority of trajectories to be diverging rather than converging. Therefore, while in some cases multiple roads may lead to the same endpoint, for the most part one would expect different roads to lead to different outcomes.

The empirical example presented below will use this benchmark of correlation with future states. As this section has made obvious, though, further thought and research on the question of OM benchmarks would be valuable.

Empirical example: occupational careers in the NLSY79

Data

The data for the empirical example come from the 1979 National Longitudinal Study of Youth (NLSY79). The NLSY79 offers detailed work history data for youth entering the labor market and follows

them into adulthood. This long-term longitudinal data is necessary for the benchmark, allowing me to measure the extent to which early occupational patterns are correlated with later occupational outcomes.

The sequence elements are the individuals' annual occupation codes at ages 21 to 30. I used the 1970 Census occupation code from the primary job at the time of each survey. I used available information to fill in missing occupation values due to missed interviews or invalid codes whenever possible. The outcome variable is occupational standing at age 35 (or 36¹²). The sample was limited to white respondents who were part of the civilian labor force (either employed with a valid occupation code or unemployed) for all ages in the study (21 to 30 and 35). The final sample was 1,013 men and 623 women.¹³

Substitution costs

I used the DOT-based approach discussed earlier to generate substitution costs for the analysis. Since the DOT includes a large number of measures, some of which assess similar concepts, I used principal component analysis (PCA) to capture the major dimensions of variation. The substitution cost matrix considers all possible pairs of occupations and calculates the sum of the differences between these pairs on the PCA components. See appendix C for details on the creation of the substitution matrix.

Alternate sequence specifications

In addition to the example presented in this paper, I conducted similar analyses using a variety of alternate approaches. For instance, I examined data from the Panel Study of Income Dynamics instead of the NLSY. I also tried expressing the sequences in terms of labor force status, major occupation groups, occupational prestige, and wages instead of detailed occupations. Finally, I also tried using transition rates to set the substitution costs in several of these cases. The results from these alternate specifications (not shown) followed similar patterns to the results presented below.

Comparing algorithms and parameter levels

The analysis compares three different methods of sequence analysis: Sequence Comparison, traditional OM, and Localized OM. Within traditional OM, I examine results at ten levels of indel costs ranging from 0.1 times the maximum substitution cost to 1.0 times the maximum cost. Within Localized OM, I tested the results at fifty different combinations of the time cost (x) and local cost (y): x = 0.1, 0.2,...1.0; y = 0.2, 0.4, 0.6, 0.8, 1.0.

Methods

The three sequence analysis techniques examined here (traditional OM, Localized OM, and Sequence Comparison) each result in dissimilarity values between every possible pair of sequences in the data, resulting in $(n * (n - 1))/2$ separate dissimilarity measures. In order to use the full extent of the output results, I use a pairwise outcome measure that is compatible with the pairwise dissimilarity results. The outcome measure, therefore, is the *difference* between the occupations held by the pair of individuals at age 35. This difference is measured as the substitution cost between the occupations. This outcome measure has the advantage that it is measured using the same metric as the original sequences. The test of each method, therefore, will be the extent to which differences between the pairs of occupational careers at ages 21-30 are correlated with differences between their occupations at age 35, with differences in occupations measured similarly throughout.

I measure the strength of the relationship between the sequence dissimilarity measures and the outcome at age 35 by estimating the following regression models:

$$w_{a_{35}b_{35}} = \beta_0 + \beta_1 D_{AB} + e \quad (8)$$

$$\begin{aligned}
w_{a_{35}b_{35}} = & \beta_0 + \beta_1 D_{AB} + \beta_2 w_{a_{21}b_{21}} + \beta_3 w_{a_{30}b_{30}} + \beta_6 w_{A_m B_m} + \beta_7 w_{A_f B_f} \\
& + \beta_4 |exp_A - exp_B| + \beta_5 |educ_A - educ_B| + \beta_7 |afqt_A - afqt_B| \\
& + \beta_8 |educ_{A_m} - educ_{B_m}| + \beta_8 |educ_{A_f} - educ_{B_f}| + e
\end{aligned} \tag{9}$$

Where:

D_{AB} = dissimilarity measure between sequences A and B using selected method

$w_{a_i b_i}$ = the substitution cost between occupations held by A and B at age i

$w_{a_m b_m}$ = the substitution cost between A and B's mothers' occupations

$w_{a_f b_f}$ = the substitution cost between A and B's fathers' occupations

exp_A = person A's years of work experience between ages 21 and 30

$educ_A$ = person A's number of years of schooling

$afqt_A$ = person A's AFQT score

$educ_{A_f}$ = person A's mother's of years of schooling

$educ_{A_m}$ = person A's father's of years of schooling

The first equation (8) measures the bivariate relationship between the dissimilarity measure (calculated using one of the three methods) and differences in occupational outcomes at age 35. The R^2 value from this regression estimate measures the strength of this relationship. The second regression equation (9) introduces a number of control variables commonly used in economic and sociological studies of employment outcomes. Since the unit of observation in these regressions is a pair of individuals, each of the control variables is expressed as a difference between the two individuals in the pair. The control variables include several occupation-based measures: occupations at age 21 and 30 (first and last year of the sequence) and mothers' and fathers' occupations when the respondents were age 14. Each of these

measures uses the substitution cost matrix to measure the difference between the two occupations. In addition, I included controls of several quantitative variables: total years of work experience, educational attainment, Armed Forces Qualifying Test score (AFQT), and parents' education levels. For these quantitative measures, I calculated the absolute value of the difference between the pair. The inclusion of all of these control variables is designed to test whether the sequence dissimilarity measures add much more to the equation than what could be captured with more commonly used measures of status attainment, family socioeconomic status, and human capital attainment.

Bootstraps

The regression analysis is based upon $n(n - 1)/2$ observation pairs even though there are only n individuals in the dataset. Given the complex nature of the data, I used bootstrapping to estimate confidence intervals for the results. In order to maintain the same structure of the data, the bootstrap samples were taken from the original list of n respondents. I then took each bootstrap sample and created a new $n(n - 1)/2$ dataset. A bootstrapped sample by definition contains duplicate observations. I was concerned that the duplicates would create pairs that compare an observation to itself, leading all of the variables in equation 9 to be zero. To avoid this problem, I replaced the pairs that compared an observation to itself with a randomly selected pair.¹⁴ The bootstrapping involved 1000 repetitions and I used the percentile technique to estimate confidence intervals; the confidence interval was the range from the 2.5th to the 97.5th percentiles of the 1000 bootstrapped results.

Results

The distribution of dissimilarities

Before I turn to the results of the regression analysis, it is useful to examine the distribution of the dissimilarity measures created by the three methods. This distribution is not a true benchmark, but it does provide some valuable insights. Figure 1 graphs the distributions of the OM dissimilarity scores using kernel density plots (smoothed histograms). The first three lines show results using the traditional

OM algorithm with the indel cost (x) set at 0.1, 0.2, and 0.3 times the maximum substitution cost. The fourth line shows the results from Localized OM. In this case I show only one parameter combination, $x = 0.1$ & $y = 0.8$. The final line shows the results from Sequence Comparison.

Figure 1 about here

The figure shows that traditional OM with $x \geq 0.3$, Sequence Comparison, and Localized OM result in distributions that are almost identical and roughly normal. I omitted the remainder of the Localized OM and higher-cost traditional OM results from the graph because they all show similar bell-shaped distributions. As one reduces the indel cost under traditional OM, however, the distributions become unusual, with patterns that reflect concerns raised earlier in the paper. Under traditional OM, a pair of indels can be used in the place of a substitution. At the extreme case, every element can be replaced using a pair of indels. Traditional OM, therefore, will never use an alignment with a dissimilarity greater than $n \times 2$ times the indel cost (20 times the indel cost in this case). As the indel cost, x , is reduced, this limit starts to cut off the right-hand tail of the distribution, as shown in the graph. The distribution for traditional OM at $x=0.1$ also has sharp peaks reflecting a reliance on indels in the place of substitutions; the peaks of the lines are located at multiples of the indel cost. In contrast to traditional OM, the distances using Localized OM are more normally distributed even at low x values as long as equation 7 holds ($y \geq 1 - 2x$), as shown by the alternate OM line for $x = 0.1$ & $y = 0.8$ in figure 1.

Regression results

Figure 2 shows one view of the pairwise regression results, providing the R^2 values from estimates of equation 8 above. One should note immediately that this figure shows an extremely narrow band of the y-axis. This presentation exaggerates the patterns I wish to highlight, the following figure will present the results at a more reasonable scale. The x-axis in figure 2 shows the values of the time cost(x) for the indels. Low values on the left side of the graph allow OM to shift sequence elements

in time at a relatively small cost. Each line on the graph represents a different setting for the local cost (y), which measures the extent to which OM punishes insertions that are different from their neighbors. As discussed previously, traditional OM is the equivalent of $y = 0$ and it is represented by a heavy solid line. A thin horizontal line on the graph shows the results for Sequence Comparison, which does not allow indels at all and therefore does not have an x -value that varies. The line is there instead to serve as a baseline against which to compare the OM results. The square markers indicate values that differed significantly from this baseline Sequence Comparison level based upon the results of the bootstrapping.¹⁵

Figure 2 about here

The figure suggests that the R^2 values are generally higher at lower indel costs. For men a number of these differences were statistically significant. For women the results showed a consistent pattern but were not statistically significant.¹⁶ Two large caveats should be added to these conclusions, however. The first caveat is that there is a limit to the improvements gained by lowering indel costs, and in fact the values drop sharply after a certain point. An inspection of the graph shows that the pattern fits equation 7: as long as y is greater than $1 - 2x$, lowering the indel cost increases the R^2 values. As soon as the equation no longer holds, though, the values drop precipitously. The peak of each line occurs at or near $y = 1 - 2x$. In other words the highest R^2 values occur when time cost (x) is set to as low as possible without allowing pairs of indels to begin replacing substitutions.

The second caveat is that the differences, while consistent particularly in the male sample, are extremely small, generally about 0.001. Figure 3 provides another perspective on the results and differs from the previous graph in a number of ways. First, the figure shows the full extent of the y -axis.¹⁷ In addition, the previous graph showed several lines for Localized OM for different levels of the local cost (y). In figure 3, I represent Localized OM with a single line that connects the points where $y = 1 - 2x$.

The results suggest that while figure 2 found that Localized OM performed better than Sequence Comparison in some cases, in figure 3 this difference is barely perceptible. The precipitous decline of traditional OM at lower indel costs, however, is still apparent.

Figure 3 about here

Finally, figure 3 presents results from regressions including the control variables described in equation 9. The figure shows the baseline R^2 value from a regression including just the control variables (the bottom line in the graph, alternating dashes and dots). The lines for Sequence Comparison, traditional OM, and Localized OM show the R^2 values when their respective dissimilarity measures are added to the regression along with the controls (R^2 values from estimates of equation 9). The gap between the control line and the sequence analysis lines, therefore, indicates the extent to which the sequence dissimilarity measures account for variation net of the other control variables. The results indicate that adding the dissimilarity measures increased the R^2 values considerably. The seven control variables together resulted in an R^2 value of 0.12 for men and 0.08 for women. The addition of the dissimilarity measures generally increased the R^2 values to about 0.15 for men and 0.12 for women. The results, therefore, suggest that the sequence analysis dissimilarity measures do seem to add something not captured by traditional economic and sociological variables.

Conclusions

The empirical example presented in this paper points to a number of important insights into Optimal Matching as a method for analyzing occupational careers. First, the results provide mixed evidence on best level of indel costs to use when examining occupational sequences. On the one hand, it appears that Abbott was correct in believing that lower-cost indels tend to be better. The R^2 values in the example tended to increase as indel costs were reduced. Abbott's suggestion of traditional OM with an indel cost of 0.1 times the maximum substitution cost, however, is a poor one because it allows too

many indels to be used in the place of substitutions and leads to very low R^2 values in the empirical example. There is, therefore, clearly a problem with traditional OM with low indel costs. The Localized OM method introduced here successfully avoids this problem, allowing the researcher to set lower indel costs without the risk of pseudo-substitutions and poor results.

The empirical example results also call into question Lesnard's view that indels are fundamentally problematic. Contrary to his assertions, OM algorithms that allowed indels performed slightly better than Sequence Comparison. Nevertheless, the improvements offered by Localized OM over Sequence Comparison were extremely small. Most of the variation captured by the three sequence analysis methods appears to have come from the comparison of sequence states, while the ability to use indels led to only a small increase in R^2 values. One has to question, therefore, whether the small improvements offered by OM are worth the additional computation involved with the method.¹⁸

The results do, however, seem to support the usefulness of sequence analysis techniques in general. With the exception of traditional OM with low indel costs, all of the methods added substantially to the explanatory power of the regression models, even after controlling for a broad range of status attainment, socioeconomic status, and human capital measures. Examining sequences as a whole, it seems, adds additional information that is not captured in other ways.

Perhaps the most important conclusion to be drawn from this study, however, is that the question of benchmarks and testing is not a trivial one. The marginal performance of OM over Sequence Comparison is a fundamental challenge to the method. OM should not be condemned, though, on the results of one study. Additional tests are necessary to ascertain whether these results are generalizable or unique to this example. In addition, the benchmark used in this paper is useful, but not perfect. Further exploration and development of appropriate benchmarks would be extremely valuable. These findings point to the need to take a step back from the ongoing applications of OM to create a better understanding of the method and how and when it is best used.

Notes

¹ One should make a distinction between the OM algorithm, which was introduced by Abbott to the social sciences, and the OPTIMIZE software program, which Abbott developed to implement the OM algorithm. Almost all social science applications of sequence analysis have used the OM *algorithm*, but not necessarily the OPTIMIZE program. In this paper I use OM to refer to the algorithm, not the software program.

² At the moment I am assuming that both sequences are of equal length n . As I will discuss later, OM is also capable of comparing pairs of sequences of different lengths.

³ See http://en.wikipedia.org/wiki/Jaccard_index and http://en.wikipedia.org/wiki/Hamming_distance

⁴ Stovel's substitution cost formula was not a simple absolute value calculation, however. She used a nonlinear formula that assigned greater differences at lower values than at higher ones (e.g. the difference between 1 and 2 was considered greater than the difference between 11 and 12).

⁵ The use of transition rates sometimes leads to confusion among people less familiar with optimal matching. They point out that transitions from s_i to s_j may be more common than transitions from s_j to s_i . Substitution costs, however, are symmetrical, requiring w_{ij} to be equal to w_{ji} . It is important to remember, though, that the use of transition rates to calculate substitution costs do not reflect an effort to actually simulate a transition but rather the idea is to use the transition rates as an indicator of distance. For example, the difference between an accountant and a laborer must be the same as the difference between a laborer and accountant. It may be easier practically to quit an accountant job and become a laborer than vice versa, but the social distance between the two positions is still the same.

⁶ Transition rates have typically been calculated using the same data source as the sequences themselves. However, one could conceivably use other datasets to make these calculations. For instance, the one-year panel structure of the Current Population Survey (CPS) in combination with its

large number of observations might allow for a measure of transition rates across detailed occupation codes. See Mouw & Kalleberg (2006) for a calculation of these types of occupational transitions.

⁷ Wu (2000: 52) notes, correctly, that with OM the dissimilarity between two sequences will be the same as the dissimilarity between the two sequences when *both* sequences are reversed (i.e. the dissimilarity between 12345 and 11223 will be the same as the dissimilarity between 54321 and 32211). This statement has been incorrectly interpreted, however, to mean that OM will find a high level of similarity between a sequence and its reverse, for instance 111000 and 000111 (Stark & Vedres 2006). These two sequences would only be considered similar to the extent that indels could be used to align the triple ones. A better example might be the sequences 12345 and 54321. These two sequences would be found quite dissimilar by OM, and in fact OM is unable to recognize situations where sequences contain common elements but in reverse order. The Elzinga and Dijkstra and Taris algorithms do have this capability, although they would “punish” these sequences for their different orders.

⁸ Consider the sequences 58941 and 05894. Assume these are scalable values, i.e. the substitution cost is the absolute value of the difference in the values. The maximum substitution cost is 9. With no indels, the difference between the two sequences is 17. If the indel cost is set at 6, well above half of 9, we can align the two sequences with an indel at the beginning and an indel at the end. There would be zero substitution costs and indel costs of 12. This alignment clearly costs less and would be chosen as the optimal alignment by OM.

⁹ In fact, biological applications sometimes use two different indel costs: a higher cost for the initial “cut” and a lower cost for indels adjacent to that cut because the fact that a cut has occurred is more important than either the values or the number of elements that are inserted into that cut.

¹⁰ Since a deletion from the perspective of one sequence is the same as an insertion from the perspective of the other sequence, this localized indel cost is easiest to comprehend by viewing all indels as insertions.

¹¹ Lesnard (2007) uses a benchmark that suffers from a similar problem. As discussed earlier, Lesnard has a strong theoretical belief that the meaning of a social phenomenon is changed when it is moved in time. His alternate algorithm, therefore, prohibits the use of indels. He applies his algorithm to empirical data and compares the results to traditional OM. His benchmark measures the extent to which individuals in the same cluster are in the same state at each time period. It comes as no surprise that Lesnard's algorithm performs better under this benchmark; traditional OM allows indels to shift sequences in time and therefore states may not necessarily match in the same time period. Lesnard's benchmark, therefore, proves that his algorithm is successful in capturing his own assumptions. The benchmark does not address, though, whether Lesnard's assumptions (i.e. that time context is of utmost importance and indels are bad) are correct.

¹² After 1994, the NLSY79 switched to biannual surveys. A number of respondents, therefore, were not interviewed at age 35, in which case I used their occupation at age 36 instead.

¹³ The sample size for women is much smaller because the sample excluded women who spent one or more years out of the labor force. An additional analysis (not shown) examined samples including individuals with periods out of the labor force. The pattern of the results was similar, but the R^2 values were lower, particularly for women. This result suggests that careers with periods out of the labor force are harder to predict.

¹⁴ Note that this replacement affected only a specific pair, not all pairs for that individual. So, for instance, if a bootstrapped sample of 1000 observations contained 150 duplicate observations, only 150 individual pairs out of the $n(n - 1)/2 = 499,500$ pairs would be replaced.

¹⁵ In each bootstrapped sample I measured the difference between the results and the R^2 value from Sequence Comparison. I then used the percentile method to estimate a confidence interval for this difference in R^2 values and the difference was significant if the confidence interval did not include zero.

¹⁶ The points on the sections of the lines in precipitous decline were statistically significant for both men and women but were beyond the scale of the graph.

¹⁷ The R^2 values presented in Figure 3 are somewhat lower than I had anticipated. Analysis using other data and sequence types yielded similar results. My preliminary exploration of this issue suggests that the low values may be an intrinsic aspect of this type of pairwise regression because all of the values are measured as differences between individuals in a pair. Similar to fixed effect models, this approach may increase the proportion of variation due to measurement error. The fact that the benchmark tests all of the alternatives under the same conditions, though, suggests that the patterns and comparative values shown here are likely valid even if measurement error is suppressing the overall results.

¹⁸ In the case of sequences with unequal lengths, however, OM may be preferable to Sequence Comparison because of its ability to stretch and compress time. The preliminary results shown here suggest that Localized OM would be a better choice than traditional OM in this case, but further testing with unequal sequences would be valuable.

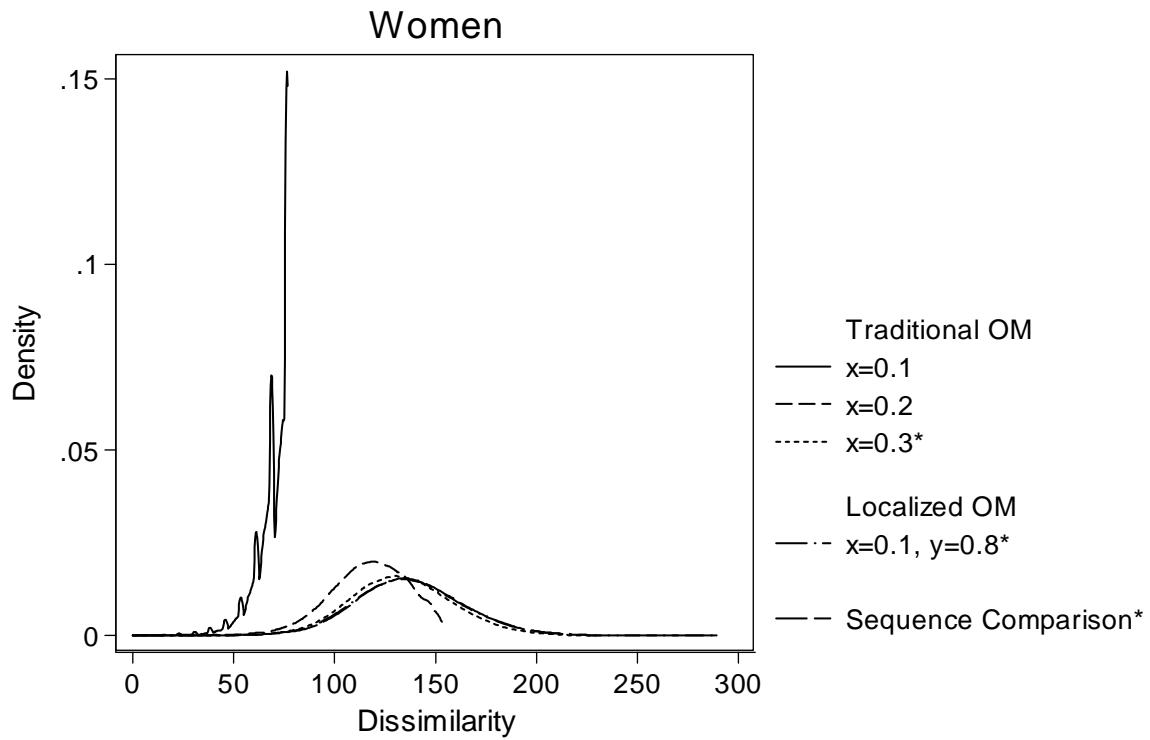
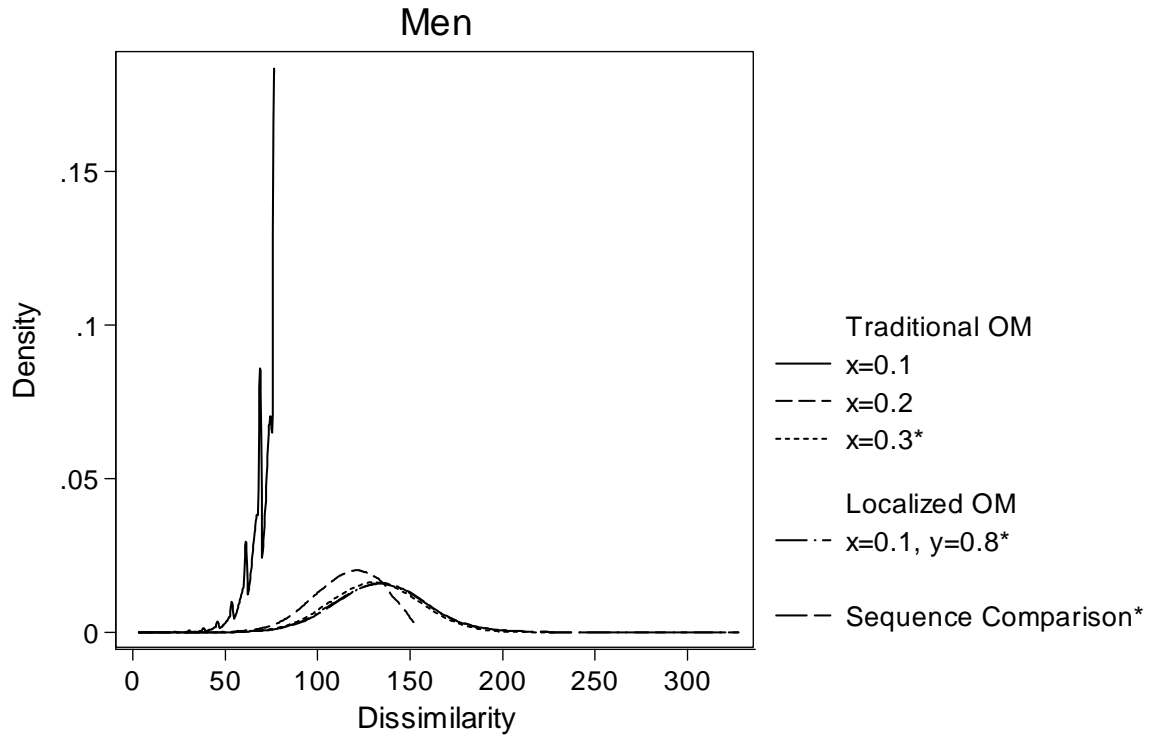
References

- Aassve, A., F. C. Billari, and R. Piccarreta. 2007. "Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories." *European Journal of Population-Revue Europeenne De Demographie* 23(3-4):369-388.
- Abbott, Andrew. 1995. "A Comment on 'Measuring the Agreement between Sequences'." *Sociological Methods & Research* 24(2):232-243.
- . 2000. "Reply to Levine and Wu." *Sociological Methods & Research* 29(1):65-76.
- Abbott, Andrew and Alexandra Hyrcak. 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musician's Careers." *American Journal of Sociology* 96(1):144-185.
- Abbott, Andrew and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect." *Sociological Methods and Research* 29(1):3-33.
- Blair-Loy, Mary. 1999. "Career Patterns of Executive Women in Finance: An Optimal Matching Analysis." *The American Journal of Sociology* 104(5):1346-1397.
- Blair-Loy, Mary and Gretchen DeHart. 2003. "Family and Career Trajectories among African American Female Attorneys." *Journal of Family Issues* 24(7):908-933.
- Bradley, David W. and Richard A. Bradley. 1999. "Application of Sequence Comparison to the Study of Bird Songs." Pp. 189-209 in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. Kruskal. Palo Alto: CSLI Publications.
- Brzinsky-Fay, Christian, Ulrich Kohler, and Magdalena Luniak. 2006. "Sequence Analysis in Stata." *The Stata Journal* 6(4):435-460.
- Clark, William A. V., Marinus C. Deurloo, and Frans M. Dieleman. 2003. "Housing Careers in the United States, 1968-93: Modelling the Sequencing of Housing States." *Urban Studies* 40(1):143.

- Dijkstra, Wil and Toon Taris. 1995. "Measuring the Agreement between Sequences." *Sociological Methods & Research* 24(2):214-231.
- Duncan, Otis Dudley. 1961. "A Socioeconomic Index for All Occupations." Pp. 109-138 in *Occupations and Social Status*, edited by A. Reiss. New York: The Free Press.
- Elzinga, Cees H. 2003. "Sequence Similarity: A Nonaligning Technique." *Sociological Methods & Research* 32(1):3-29.
- 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.
- Han, Shin-Kap and Phyllis Moen. 1999. "Clocking Out: Temporal Patterning of Retirement." *American Journal of Sociology* 105(1):191-236.
- Harding, David J. 2007. "Cultural Context, Sexual Behavior, and Romantic Relationships in Disadvantaged Neighborhoods." *American Sociological Review* 72(3):341-364.
- Huang, Q. H. and M. Sverke. 2007. "Women's Occupational Career Patterns over 27 Years: Relations to Family of Origin, Life Careers, and Wellness." *Journal of Vocational Behavior* 70(2):369-397.
- Keister, Lisa A. 2003. "Religion and Wealth: The Role of Religious Affiliation and Participation in Early Adult Asset Accumulation." *Social Forces* 82(1):175-207.
- Kogan, I. 2007. "A Study of Immigrants' Employment Careers in West Germany Using the Sequence Analysis Technique." *Social Science Research* 36(2):491-511.
- Lesnard, Laurent. 2007. "Describing Social Rhythms with Optimal Matching." French National Center for Scientific Research, Observatoire Sociologique du Changement Unpublished Working Paper.
- Levine, Joel H. 2000. "But What Have You Done for Us Lately?: Commentary on Abbott and Tsay." *Sociological Methods & Research* 29(1):34-40.

- MacIndoe, Heather and Andrew Abbott. 2004. "Sequence Analysis and Optimal Matching Techniques for Social Science Data." Pp. 387-406 in *Handbook of Data Analysis*, edited by M. A. Hardy and A. Bryman. Thousand Oaks: Sage.
- McVicar, Duncan and Michael Anyadike-Danes. 2002. "Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 165(2):317-334.
- Mouw, Ted and Arne L. Kalleberg. 2006. "Intragenerational Mobility, Careers, and Task-Specific Human Capital." in *American Sociological Association Annual Meeting*. Philadelphia, PA.
- National Academy of Sciences. 2001. "Dictionary of Occupational Titles (DOT): Part I- Current Population Survey, April 1971 Augmented with DOT Characteristics." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Washington, DC: U.S. Dept of Commerce, Bureau of the Census [producer].
- Pollock, Gary. 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 (Statistics in Society)* 170(1):167-183.
- Pollock, Gary, Valerie Antcliff, and Rob Ralphs. 2002. "Work Orders: Analysing Employment Histories Using Sequence Data." *International Journal of Social Research Methodology* 5(2):91-105.
- Sankoff, David and Joseph Kruskal. 1999. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Palo Alto: CSLI Publications.
- Scherer, Stefani. 2001. "Early Career Patterns: A Comparison of Great Britain and West Germany." *European Sociological Review* 17(2):119-144.
- Stark, David and Balázs Vedres. 2006. "Social Times of Network Spaces: Network Sequences and Foreign Investment in Hungary¹." *The American Journal of Sociology* 111(5):1367.

- Stovel, Katherine. 2001. "Local Sequential Patterns: The Structure of Lynching in the Deep South, 1882-1930." *Social Forces* 79(3):843-880.
- Stovel, Katherine and Marc Bolan. 2004. "Residential Trajectories: Using Optimal Alignment to Reveal the Structure of Residential Mobility." *Sociological Methods & Research* 32(4):559-598.
- Stovel, Katherine, Michael Savage, and Peter Bearman. 1996. "Ascription into Achievement: Models of Careers Systems at Lloyds Bank, 1890-1970." *American Journal of Sociology* 102(2):358-399.
- Vingron, Martin and Michael S. Waterman. 1994. "Sequence Alignment and Penalty Choice: Review of Concepts, Case Studies, and Implications." *Journal of Molecular Biology* 235:1-12.
- Williams, Sonya and Shin-Kap Han. 2003. "Career Clocks: Forked Roads." Pp. 80-97 in *It's About Time: Couples and Careers*, edited by P. Moen. Ithaca, NY: Cornell University Press.
- Wu, Lawrence. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect'." *Sociological Methods & Research* 29(1):41-64.



*Note: the three lines marked by * are so similar that they are difficult to discern in these graphs.

Figure 1. Distribution of dissimilarity values.

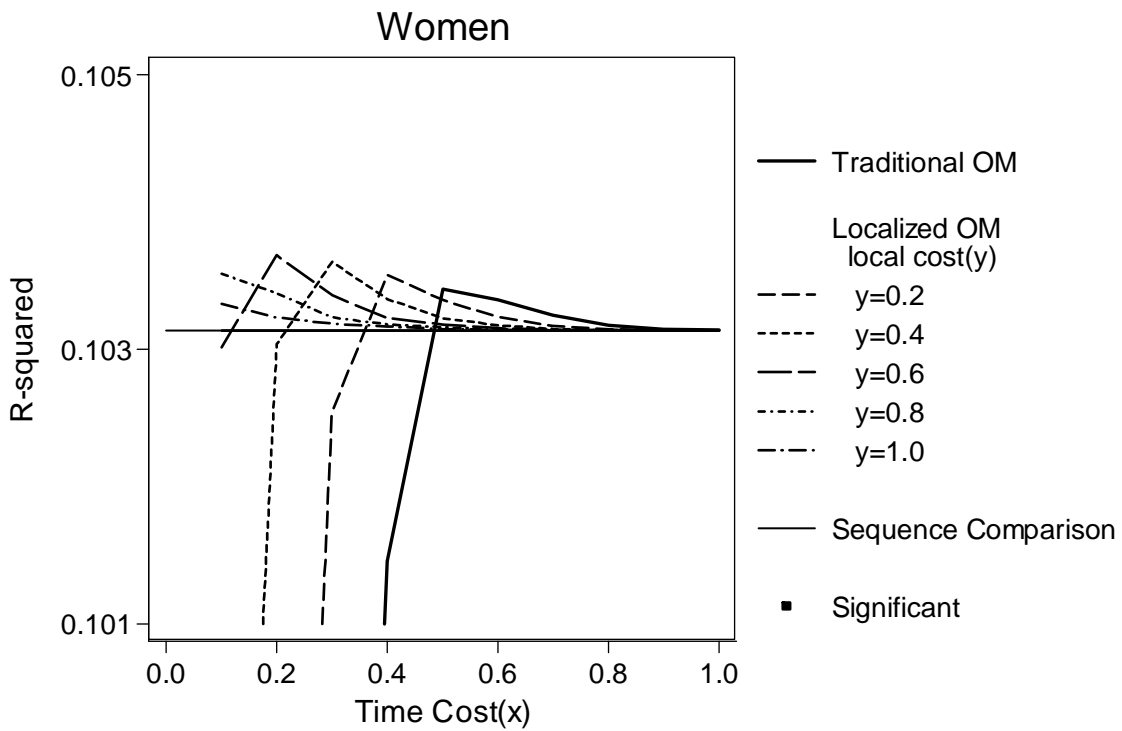
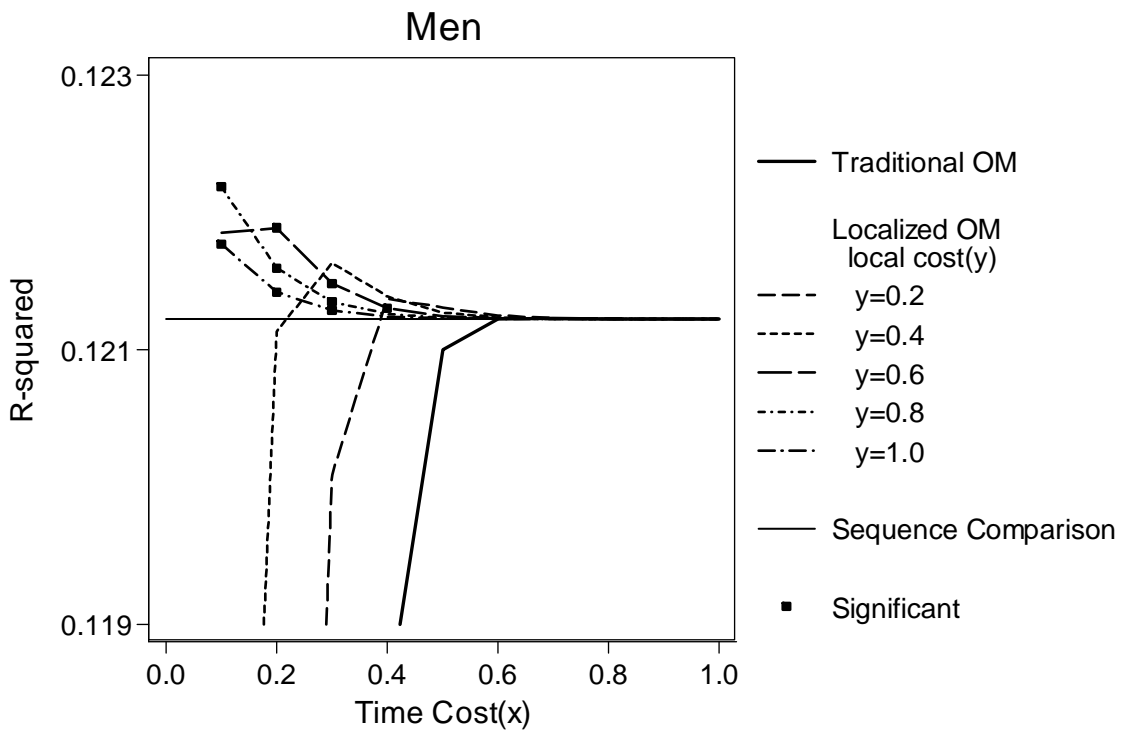
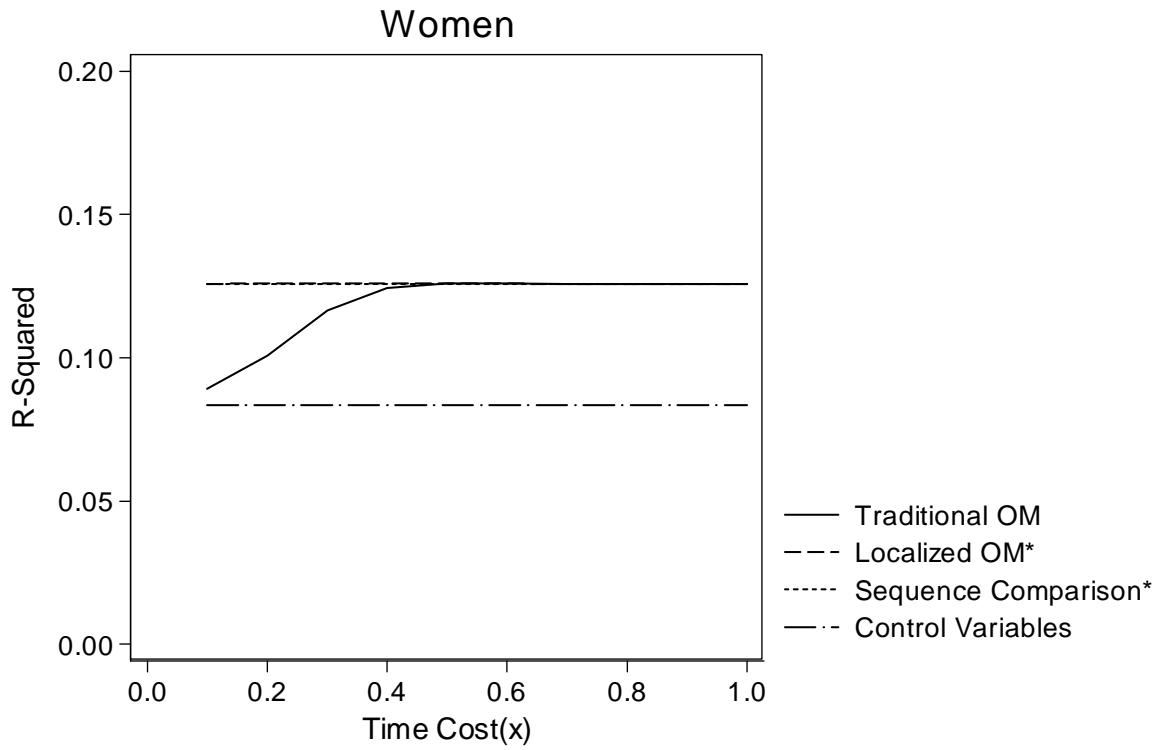
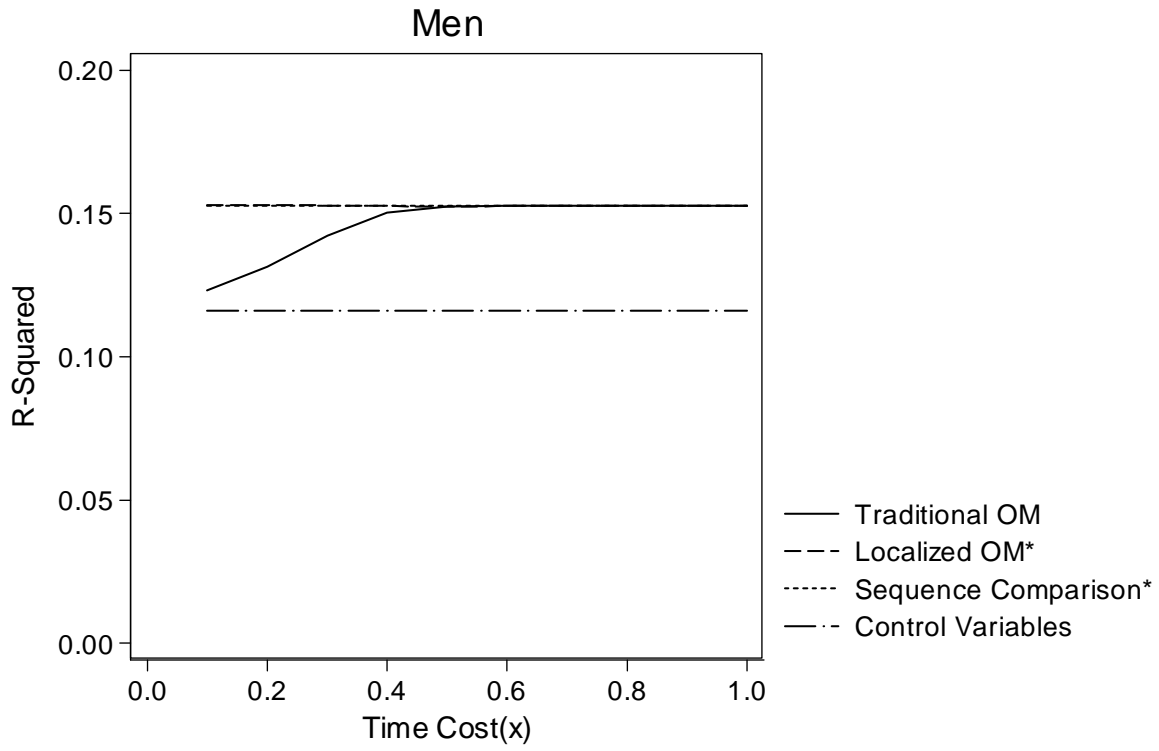


Figure 2. R-squared values from pairwise regression, zoomed view.



*note: the difference between Localized OM and Sequence Comparison is almost imperceptible in these graphs.

Figure 3. R-squared values from pairwise regression including control variables.

Appendix A: Sequences of unequal lengths

One of the potential strengths of optimal matching is that it allows the researcher to compare sequences of different lengths. The best way to set indel costs for sequences of unequal lengths, though, is a question that has not been resolved. One issue is that there are two possible causes of unequal length sequences: actual differences in sequence durations and data censoring. These two situations call for two different approaches to indels.

The first case involves sequences that differ in duration for substantive reasons. Stovel et al (1996) found this situation in their analysis of careers at a specific bank; some careers at the bank were longer than others. In this situation they chose not to make any specific accommodations, the OM algorithm will “stretch” the shorter sequence to make it the same length as the longer sequence by using insertions at optimal locations. The insertions used to equalize the sequence lengths will add to the overall “cost,” which is appropriate because the fact that the lengths are different should be a source of dissimilarity. This approach could be used with either traditional or Localized OM. The time cost (x) of an indel determines how much to “charge” for stretching or compressing time, which would similarly apply to the indels required to make the lengths comparable. One issue to consider, though, is that with sequences of equal length indels always have to come in pairs. If one inserts an element into a sequence, then one has to delete an element somewhere else (or insert an element in the second sequence, which is in effect the same thing). The indels required with sequences of unequal length, however, do not have to come in pairs. One might want to consider, therefore, charging extra, or perhaps even double, the normal indel cost for these shifts.¹

¹ Stovel and Bolan (2004) used varying indel costs for sequences of unequal lengths, although for a case of censored sequences, which will be discussed in a moment. They chose to set *lower* indel costs for unequal sequences. Their decision, however, was primarily based upon concerns that would actually have been solved with Localized OM (see the discussion in their appendix). They also chose to set a lower cost for *all* indels used when

A second potential cause of unequal lengths is right- or left- censoring of the data. The most obvious cause of this censoring is a limited window of observation. In the case of censored sequences, I would argue for looking only at the common observations. So, for instance, if sequence A is left-censored and only has observations of the fifth through tenth years while the sequence B has data for all ten years, I would argue for looking only at the commonly observed elements (years five through ten) in the two sequences. Other approaches might lead to systematic biases because optimal matching would be aligning based upon more complete information on one sequence versus the other. The challenge with this approach, though, is that the resulting dissimilarity values between pairs would be dependent upon the number of elements compared in the pair. Measuring the average dissimilarity per element compared might address this situation.

comparing unequal sequences. I would argue, however, that if one is going to use different indel levels for unequal sequence, this different level should be only applied to the indels specifically required to get the sequences to equal lengths, not to all indels.

Appendix B: Derivation of x & y parameter formula

Consider the following two sequences.

Sequence 1: a a a a a a

Sequence 2: a a a t a a

Assume the states a and t are maximally different ($w_{at} = w_{max}$). The OM algorithm can get around this substitution by shifting some of the elements over and inserting a t in sequence 2 in the proper place.

Since sequence 2 would then have an extra element, one of the a 's would have to be deleted (or, from another perspective, one would have to insert an extra a into sequence 1). With the localized OM

approach, the cheapest place to do this would be between two other a 's (minimizing the local cost). The goal is to make this combination of two indels cost more than the largest substitution cost, so that the

OM algorithm will never choose to do this "pseudo-substitution."

$$w_{max} \leq C_{ata} + C_{aaa} \quad (10)$$

Equation 6 in the paper showed us that:

$$C_{izj} = x \cdot w_{max} + y \cdot \frac{(w_{zi} + w_{zj})}{2} \quad (6)$$

When inserting t between two a 's, w_{zi} and w_{zj} are both equal w_{max} :

$$C_{ata} = x \cdot w_{max} + y \cdot \frac{(w_{max} + w_{max})}{2} \quad (11)$$

For the second indel, which is equivalent to inserting an a between two other a 's, w_{zi} and w_{zj} are both zero:

$$\begin{aligned}
C_{aaa} &= x \cdot w_{max} + y \cdot \frac{(w_{aa} + w_{aa})}{2} \\
&= x \cdot w_{max} + y \cdot \frac{(0 + 0)}{2} \\
&= x \cdot w_{max}
\end{aligned}
\tag{12}$$

Based on the criteria in equation 10:

$$w_{max} \leq x \cdot w_{max} + x \cdot w_{max} + y \cdot \frac{(w_{max} + w_{max})}{2}
\tag{13}$$

Which reduces to:

$$1 - 2x \leq y
\tag{7}$$

As long as equation 7 holds, a pair of indels will not be used to replace a substitution.

Appendix C: DOT-based substitution costs

The National Academy of Sciences (2001) provides a dataset of the April 1971 Current Population Survey with Dictionary of Occupational Titles (DOT) measures added to each record based upon the original description of the respondent's job (therefore allowing more detailed occupation coding than the census occupation codes). The DOT provides a large number of measures, some of which capture similar concepts (see the dataset codebook <http://dx.doi.org/10.3886/ICPSR07845> for a detailed description). In order to reduce the redundancy of the measures, I used principal component analysis (PCA) on the DOT variables,² retaining the first ten components of the PCA results. I then grouped the individual records and calculated the average of each PCA component within each of the 1970 census occupation codes.³

The substitution cost matrix was created by comparing each pair of occupations, taking the absolute value of the differences between this pair on each of the PCA components, and then adding these differences together.

$$w_{ij} = \sum_{k=1}^{10} |pca_{ki} - pca_{kj}| \quad (14)$$

where:

w_{ij} = substitution cost between occupation i and occupation j

pca_{ki} = value of PCA component k for occupation i

Since the PCA components are orthogonal to each other, this substitution cost matrix is multi-dimensional. The PCA components are created such that the scale of each component is reflective of the

² I chose not to include the “people, places, and things” measures because the DOT warns researchers against treating these variables as ordinal.

³ The calculation was a weighted average using CPS-provided survey weights.

amount of variation captured by that component (i.e. the first component has the greatest standard deviation). Each PCA component, therefore, is naturally weighted to reflect its overall contribution to variation across occupations in work characteristics.

The DOT does not provide measures for unemployment. I therefore set the state of unemployment to have the minimal value on all of the DOT measures (lowest skill levels, absence of any characteristics, etc) and included it in the calculation of substitution costs.

Appendix D: Previous studies using OM

Citation	Topic	Substitution costs	Indel cost/ w_{\max}	Use of results
Abbott & Hrycak (1990)	Careers of German musicians, 1950-1810	Combination of transition rates and theory	≥ 1	Cluster analysis using multiple methods to identify robust patterns
Stovel, Savage & Bearman (1996)	Careers within Lloyd's bank 1890-1970	Transition rates	1	Identification of common clusters separately for three time periods, comparison of resulting clusters in each time period
Halpin & Chan (1998)	Intragenerational class mobility	Theoretically derived	0.75 "set relatively high"	Multidimensional scaling to examine coherence of patterns, cross-cohort differences in cluster membership
Blair-Loy (1999)	Careers of women in finance-related jobs	Combination of transition rates and theoretical	0.48, based on experimentation	Cluster analysis, factors behind patterns, measured whether dissimilarity within clusters decreased over time
Han & Moen (1999)	The retirement process	Transition rates	Not specified	Clusters used as dependent variables predicting a number of retirement behaviors.
Scherer (2001)	Career patterns	Theoretically derived	0.5	Deviation of work histories from ideal-type of full-time work, comparison of Britain and Germany in cluster frequencies
Stovel (2001)	Local lynching histories in the Deep south 1882-1930	Scalar substitution costs with non-linear transformation	Value of inserted/deleted element (b/c scalar)	Identification of common clusters
McVicar & Anyadike-Danes (2002)	School to work transitions	Theoretically derived	0.5, also tested 1.33	Clusters are used as dependent variables in logit estimates to identify background characteristics most influential in determining path

Citation	Topic	Substitution costs	Indel cost/w_{\max}	Use of results
Pollock, Antcliff & Ralphs (2002)	Work careers	Adapted from transition rates	0.5	Identification of common clusters, cluster membership by sex
Blair-Loy & DeHart (2003)	Career histories of African-American female lawyers	Not specified	Not specified	Used a subset of clusters as dependent variables in logit estimates to identify factors predicting career paths.
Clark, Duerloo & Dieleman (2003)	Housing careers	Theoretically derived	>1	Started with a preset list of 26 basic types, used OM to classify into these types. Compared frequencies by age of household head, income, and region.
Keister (2003)	Wealth patterns	Transition rates	Not specified	Identification of common clusters, comparison of cluster frequencies by religious affiliation
Williams & Han (2003)	Multiple dimensions of work careers	Not specified	Not specified	Averages within clusters of sorting mechanisms (gender, education, birth cohort) and outcomes (salaries, perceived success & security, family events, etc)
Arosio (2004)	Occupational careers	Theoretically derived	0.71, not explained	Identification of common clusters
Keister (2004)	Wealth patterns	Transition rates	Not specified	Identification of common clusters, comparison of cluster frequencies by family background characteristics
Stovel & Bolan (2004)	Residential trajectories	Theoretically derived adjusted with transition rates	>1, with lower indel costs for sequences of unequal lengths	Identification of common clusters in subsamples, classification of full sample using representative sequences
Stark & Vedres (2006)	Evolution of firm networks	Transition rates	>1 based on empirical tests	Identification of common clusters, relationship to foreign ownership

Citation	Topic	Substitution costs	Indel cost/w_{max}	Use of results
Harding (2007)	Sequencing of events in romantic relationships	Transition rates	1, cites Stovel et al. (1996)	Dissimilarity between ideal and actual romantic relationship is used as a dependent variable, amount of variation at the neighborhood level in these sequences used as an independent variable
Lesnard (2007)	Daily work shift patterns	Time-variant transition rates	Indels not allowed	Compared homogeneity of clusters using new method, unit cost OM, and variable substitution OM
Pollock (2007)	Combined employment, housing and family careers	Transition rates	0.5	Identification of common clusters
Kogan (2007)	Comparison of native and immigrant employment careers in Germany	Unit costs	0.5	Dissimilarity from a sequence of employment in all periods, identification of common clusters
Huang & Sverke (2007)	Women's occupational careers	Theoretically derived	Lower than 0.5	Identification of common clusters, relationship to family of origin as well as quality of life
Aassve, Billari & Piccarreta (2007)	Work-family trajectories of young British women	Transition rates	>0.5	Identification of common clusters, mediod sequences as descriptions of clusters, graphical presentation of clusters