Optimal matching analysis in career research: A review and some best-practice recommendations

Katja Dlouhy *, Torsten Biemann

Department of Management, University of Mannheim, Mannheim 68161, Germany

ARTICLE INFO

ABSTRACT

Optimal matching is a method for the analysis of sequential data. It allows researchers to detect patterns in career sequences or occupational trajectories. We first give a brief introduction to the method and review the existing career literature that employs optimal matching. To examine which data properties are required for optimal matching analysis, we conducted Monte Carlo simulations of career sequences with varying parameters for sequence length, sample size and missing items. We find that sequence length is the relevant factor for correct results, while sample size does not substantially affect result quality. Another important finding is that sequences with up to 30% elements missing can be used for optimal matching analysis. We also show which settings for the optimal matching procedure deliver the best results.

1. Introduction

A career can be viewed as an unfolding sequence of a person's work experience over time (Arthur & Rousseau, 1996). Career studies or studies of vocational development examine differences within an individual at two or more points in time, and longitudinal studies concerning career patterns are therefore needed (Savickas, 2002). The definition by Super (1980: 282) of a career as a "sequence of positions occupied by a person during the course of a lifetime" emphasizes the longitudinal character important in career research. Occupational trajectories also depict occupational statuses of an individual in a certain time span, but a career refers to reflections about the course of changes in vocational behavior over time (Savickas, 2002). When studying careers and occupational trajectories, it thus makes sense to study sequences of occupational states as wholes.

Numerous sequence analysis methods have been developed over the last decades, and optimal matching analysis (OMA) is the method that has received by far the most attention in the social sciences (Aisenbrey & Fasang, 2010). OMA is a quantitative method that was mainly used in the natural sciences to find resemblances between proteins or DNA sequences, before it migrated to the social sciences.

With OMA, researchers can assess the similarity of sequences by assigning costs to operations that would be needed to align these sequences. Sequences are then clustered according to similarity; sequences in one cluster are assumed to follow the same career pattern. An advantage of OMA over other clustering methods is that it takes in sequences as wholes. The focus is not on single states in specific points in time, but on holistic trajectories. Thus, optimal matching is a convenient method for the analysis of careers and occupational trajectories. Findings from OMA can be used to advance theory on vocational development processes. OMA has often been used merely for descriptive studies, primarily in sociological studies. However, OMA results can also be used for explanatory purposes. Belonging to certain career sequence clusters or following specific occupational trajectories can be linked to other variables, like work attitudes (Huang & Sverke, 2007) or career success (Kovalenko & Mortelmans, 2014).

---

* Corresponding author at: University of Mannheim, Department of Management, Mannheim 68161, Germany.
E-mail address: katja.dlouhy@bwl.uni-mannheim.de (K. Dlouhy).

Researchers can use existing longitudinal data or generate new data using retrospective collection methods, but the data should in any case be appropriate for OMA. However, there is no precise information to be found that answers some very basic, but important questions. Whether there is a minimum sample size or sequence length, which data should have for OMA, has not been subject to any study yet. The effects of missing elements or inequality in sequence length on OMA results are also unclear. Likewise, questions regarding which cost settings should be used or which clustering algorithms deliver the best OMA results are answered ambiguously in the existing literature. We therefore simulate career data with different parameters. The results of these studies enable us to provide guidelines for researchers. By giving recommendations on sample size, sequence length, and maximum number of missing items that are required for OMA, we hope to facilitate the use of this method in career research.

The paper is organized as follows. After giving a brief introduction to the method, we review studies on careers and occupational trajectories that employed OMA so far. Our focus is not only on study outcomes, but also on properties of the data that were used. We then describe how we simulated data with different sample sizes, sequence lengths or numbers of missing items. Subsequently, we show how the quality of the OMAs' results changes when specific cost settings and clustering algorithms are used. The results of the simulation studies enable us to provide best-practice recommendations on the use of OMA. Finally, we discuss the results and possible limitations of our studies and show possibilities for future research.

2. Optimal Matching Analysis

In this section we briefly explain the logic of optimal matching. Optimal matching can be used to analyze sequences that consist of elements with a certain spatial ordering, like nucleotide sequences in DNA strands, or a certain temporal ordering, like career states in subsequent years. When conducting an OMA of sequential data, the elements in the sequences first need to be coded with a finite amount of numbers or letters. For an example of occupational trajectory sequences which are coded with employment states, see Table 1.

Similarity of a pair of sequences can be assessed by calculating metric distances between them. Metric distances are created by setting costs for operations which are needed to transform one sequence into another. The more operations are needed, the higher are the costs and the less similar are the sequences. To align the sequences in our example, we use the following operations: insertion of an element, deletion of an element, and substitution of one element for another. We choose a setting where insertion and deletion (hereafter shortly referred to as indel) operations cost 1 and substitution costs 2. The absolute height and the measurement unit for the costs are irrelevant (Biemann & Datta, 2014), only the relationship between substitution and indel costs is important (Aisenbrey & Fasang, 2010).

Costs for one substitution should not be higher than the costs for one insertion and one deletion. Otherwise substitutions would not be used. In our example, substitution operations can be used interchangeably with one insertion and one deletion operation. If substitution costs are set lower than the costs for one deletion and one insertion, this should be based on theoretical considerations. Substitution costs can also be based on transition rates from one state to other states that are derived directly from an existing data set (e.g., Biemann, Zacher, & Feldman, 2012). The rationale for using a transition-based substitution cost matrix is that the frequency of transitions between states gives information about the extent of similarity between these states (Biemann & Datta, 2014; Stark & Vedres, 2006).

We want to align sequence A (*EEEEEUEE*) and sequence B (*EEEEEEE?*). There are two differences between those sequences. In the sixth year, individual A was unemployed, whereas individual B was employed. In the eighth year, individual A was employed, whereas for individual B, there is no information on employment status. When transforming sequence A into sequence B, the cost is 1 for deleting the *U* from sequence A. After deletion of this element, both sequences consist of seven states *E* and one missing state at the end, and are thus equal.

The optimal matching algorithm can thus warp sequences' timely ordering. This is possible because only patterns in sequences are of interest. Elements can be inserted everywhere, so that sequences with missing elements can also be analyzed. The lower the cost (or metric distance), the more similar are the sequences. What should be noted here is that the costs for transforming sequence A into sequence B are the same as for transforming sequence B into sequence A, and that transforming a sequence into itself has zero costs.

When transforming sequence C (*EEUUSSSS*) into sequence D (*EEUUUSSS*), we could either delete the *S* in the fifth year in sequence C, which costs 1, and insert an *U* instead, which also costs 1. These operations would have an overall cost of 2. We could also substitute the first *S* in sequence C for an *U*, which would also cost 2. There are multiple ways of transforming one sequence into another. The optimal matching algorithm always calculates the lowest possible metric distance between sequences. After calculating metric distances for each pair of sequences, we get a resulting matrix (see Table 2).

Table 1
Example of occupational trajectories.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8
Trajectory A	E	E	E	E	E	U	E	E
Trajectory B	E	E	E	E	E	E	E	?
Trajectory C	E	E	U	U	S	S	S	S
Trajectory D	E	E	U	U	U	S	S	S

Note. E = Employed, S = Self-employed, U = Unemployed, ? = Missing.