Evaluating the Effects of Missing Values and Mixed Data Types on Social Sequence Clustering Using t-SNE Visualization

Alina Lazar[1], Ling Jin[2], C. Anna Spurlock[2], K. John Wu[2], Alex Sim[2], Annika Todd[2]

[1] Youngstown State University, Youngstown, OH, USA
[2] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Disclaimers
This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Evaluating the Effects of Missing Values and Mixed Data Types on Social Sequence Clustering Using t-SNE Visualization

*

ALINA LAZAR[*], Youngstown State University, USA
LING JIN[*] and C. ANNA SPURLOCK, Lawrence Berkeley National Laboratory, USA
KESHENG WU and ALEX SIM, Lawrence Berkeley National Laboratory, USA
ANNIKA TODD, Lawrence Berkeley National Laboratory, USA

The goal of this paper is to investigate the impact of missing values in clustering joint categorical social sequences. Identifying patterns in socio-demographic longitudinal data is important in a number of social science settings. However, performing analytical operations, such as clustering on life course trajectories, is challenging due to the categorical and multi-dimensional nature of the data, their mixed data types, and corruption by missing and inconsistent values. Data quality issues were investigated previously on single variable sequences. To understand their effects on multivariate sequence analysis, we employ a dataset of mixed data types and missing values, a dissimilarity measure designed for joint categorical sequence data, together with dimensionality reduction methodologies in a systematic design of sequence clustering experiments. Given the categorical nature of our data, we employ an ”edit” distance using Optimal Matching (OM). Because each data record has multiple variables of different types, we investigate the impact of mixing these variables in a single dissimilarity measure. Between variables with binary values and those with multiple nominal values, we find that the ability to overcome missing data problems is more difficult in the nominal domain than in the binary domain. Additionally, alignment of leading missing values can result in systematic biases in dissimilarity matrices and subsequently introduce artificial clusters as well as unrealistic interpretations of associated data domains. We demonstrate the usage of t-distributed Stochastic Neighborhood Embedding (t-SNE) to visually guide mitigation of such biases by tuning the missing value substitution cost parameter or determining an optimal sequence span.

CCS Concepts: • **Computing methodologies → Dimensionality reduction and manifold learning**; *Cluster analysis*; *Feature selection*;

Additional Key Words and Phrases: Joint sequence analysis, optimal matching, missing values, time series clustering, data quality, t-SNE, dimensionality reduction, life trajectories.

[*] Alina Lazar and Ling Jin contributed equally to the work.

Authors' addresses: Alina Lazar[*], Department of Computer Science and Information Systems, Youngstown State University, One University Plaza, Youngstown, OH, 44555, USA, alazar@ysu.edu; Ling Jin[*]; C. Anna Spurlock, Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory, Berkeley, USA, ljin@lbl.gov; Kesheng Wu; Alex Sim, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA, USA; Annika Todd, Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory, Berkeley, USA.

1936-1955/2019/3-ART7 $15.00
https://doi.org/0000001.0000001

ACM Reference Format:
Alina Lazar*, Ling Jin*, C. Anna Spurlock, Kesheng Wu, Alex Sim, and Annika Todd. 2019. Evaluating the Effects of Missing Values and Mixed Data Types on Social Sequence Clustering Using t-SNE Visualization: . *ACM J. Data Inform. Quality* 11, 2, Article 7 (March 2019), 28 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Time series clustering plays an important role in temporal data mining research [Kotsakos et al. 2013]. In this work, we study the quality issues often present in many common sources of time series data. To make this exploration concrete, we use the task of clustering multivariate life course trajectories consisting of mixed data types from a large collection of real world survey data with prominent and not commonly addressed data quality issues. We assess the effects of data quality issues by comparing and diagnosing clustering solutions from a systematically designed set of dissimilarity measures using internal validity metrics, normalized mutual information, and dimensionality reduction techniques. This task reveals the extent to which a number of data quality issues such as missing values, data consistency issues, and mixed data types make it challenging to compare time series sequences. For example, how the missing values are handled could significantly affect or bias the ”dissimilarity” measures and therefore change the clusters derived. We see value to this investigation across a variety of applications. For example we anticipate that similar challenges are present in analysis of other trajectory data, such as sensor data where sensor malfunction could be common especially in large-scale deployments and for continuous data collection. Other possible applications include characterizing market segments from customers' purchasing history data for the purposes of targeted advertising, identifying symptom triggers for asthma patients using data collected through a mobile health applications, hourly electricity data where readings are often missing and customers may have different time ranges, or other similar long-term tracking and categorization exercises using real world data.

In social science research, time series clustering is used to study the spans of individuals' life trajectories in the form of sequence analysis [Bras et al. 2010; Schumacher et al. 2012; Widmer and Ritschard 2009] . By analyzing long-term life trajectory dynamics based on demographic characteristics, education and other lifestyle variables, it is possible to discover representative patterns within a population based on the overall life trajectory of a given individual's characteristics and the pathway through which one arrives at a given state, decision or behavior. Given the many factors and their interdependences that affect life trajectories, such as family planning, education, or employment, joint sequence analysis (formalized by [Gauthier et al. 2010; Pollock 2007]) often represents a more appropriate method than single-variable sequence analysis. Its power relies on its capability to differentiate longitudinal experiences represented by multiple variables and therefore accounts more realistically for the inherent complexity of life trajectory patterns [Johnson and Onwuegbuzie 2004; Lauder et al. 2004; Wiles 2004].

One challenge in clustering life course trajectories with most common sources of such data is the missing data problem. Discarding sequences with missing values comes with the sacrifice of losing sample representativeness. Despite the benefit of providing a contextual and dynamic view of individuals' life courses, most panel data sources are especially prone to missing values, as missing data often arise from the difficulty in repeated collection of data on a continuous and consistent basis for each individual. This is a particularly common problem in the context of panel surveys where the same individuals are contacted repeatedly over long-term time horizons. This type of missing data will be referred to as *survey gaps.* Another type of missing value, even with otherwise ”perfect” longitudinal survey with complete data records for individuals during the survey period, arises from sequence alignment by development time (e.g., age) instead of calendar time. When the sequences are aligned by age, different age cohorts will appear with unequal sequence lengths.

Individuals that enter the data collection at an older age will inevitably miss the leading segment of their life course data. Censored data have similar contiguous missing observations at the beginning and/or ending of the sequences, which can be due to reasons other than alignment by age. For simplicity, this type of missing data will be referred to as *alignment missing*. Both types of missing values (survey gaps and alignment missing) are often encountered in real-world data, however their effects are not adequately evaluated especially in the joint sequence analysis literature.

Another challenge in clustering life course trajectories is the mixed data type problem in joint sequence analysis. Categorical variable types with different numbers of state spaces (e.g. binary variables versus nominal variables with multiple states) affect their contributions to the distance measures determined in the joint domain. Consequently, clustering derived from the joint domain may have different representation of individual variables and associated cluster interpretations as illustrated in [Piccarreta 2017]. As survey gaps or alignment missing usually affect multiple variables the same way, missing values may complicate the association among variables considered in the joint sequences and lead to potentially incorrect interpretations. Such interactive effects from both missing values and mixed data types have yet to be examined in the literature.

Lastly, in joint sequence analyses, the data dimensions increase as the number of time dimensions begins to multiply with the number of variables included, leading to difficulties in exploration and diagnosis of the clustering results in their original sequence representation. Dimensionality reduction is an essential tool to capture the qualitative cluster structure in a low dimensional space that can be easily visualized when working with the high dimensionality aspect of the ”largenes” of data. Limited usage of such techniques is present in the social sequence analysis literature especially with the application of non-linear dimension reduction techniques.

The goal of this paper is to provide a systematic investigation of the aforementioned two challenges, missing values and mixed data types, in joint sequence analysis. This study contributes to the social sequence analysis literature by: (1) including an under-studied type of missing value that arise from data alignment, where imputation is often not practical; (2) assessing missing value problems present in multiple-variable as opposed to single-variable sequence analysis; (3) highlighting the interactions between the challenges of both missing value treatments and mixed data types; (4) demonstrating the usage of nonlinear dimensionality reduction techniques for social sequence analysis applications.

The rest of the paper is organized as follows. Section 2 provides a survey of related literature. Section 3 describes the real-world data we use for the study. Section 4 explains the methods we use, including the distance measures, the systematic experimental clustering design, the comparison metrics we use, and dimensionality reduction techniques used for visualization and diagnosis. Section 5 evaluates clustering results, their dependence on missing value treatments and mixed data types, and bias mitigation methods. Section 6 concludes.

2 RELATED WORK

The notion of clustering hinges on the notion of distance, and therefore the concept and quantification of dissimilarity is important for time-series data clustering. Metric distances such as the Euclidean distance fit well as dissimilarity measures and are applied widely to identify patterns in longitudinal numeric data [Fu 2011; Jin et al. 2017; Liao 2005; Rani and Sikka 2012]. Given the categorical (as opposed to continuous numeric), longitudinal characteristics of the life trajectory sequences commonly encountered in social sciences, the classical clustering approach based on metric distances does not work well. Since its introduction by Andrew Abbott [Abbott and Forrest 1986; Abbott and Hrycak 1990] the *edit* based dissimilarity measure through Optimal Matching (OM) has become the most common way of computing dissimilarities between sequences describing life trajectories of multiple individuals.

OM was used first in molecular biology for comparing and analyzing DNA sequences [Needleman and Wunsch 1970] and also in natural language processing for approximate character string matching [Wagner and Fischer 1974]. The OM method uses counts of sequence alignment operations such as inserts, deletions (*indels*) and substitutions to transform one sequence to resemble another one. The fewer steps needed for the transformation, the closer the two sequences are considered. The final distance value calculated based on the OM procedure is deeply affected by the costs assigned to the indel and substitution transformations. Different setups and combinations [Studer and Ritschard 2016] proposed for specific problems hold various results. Adjusting the transformation cost allows researchers to tune if the algorithm will use more or fewer indels or substitutions. Sequence similarity can be described in terms of when elements occur or in terms of the order of these elements. When order is more important than timing it is recommended to reduce the number of substitutions by increasing their cost while keeping the indels cost equal to 1. Setting substitution cost to the same value of 1 or 2 assumes that all these substitution transformations are equally important. However, some transitions between states are more realistic therefore computing the substitution cost based on each transition probability would be better. More research is required to fully understand the consequences of different cost regimes. Also, assigning substitution costs for transitions to the missing states has not been investigated.

Traditional sequence analysis as well as new method development or evaluation have mostly focused on single variable trajectories [Elzinga 2007; Gauthier et al. 2009; Halpin 2014; Studer and Ritschard 2016; Studer et al. 2011]. This approach allows for relatively easy evaluation of missing observations that affects many real-world data sequences. For aforementioned types of data missing, it is relatively straightforward to address the short internal survey gaps as they can be imputed using the before and after data present in the sequence. Evaluation of the imputation strategy of these internal gaps has been the focus of a small number of past studies [Halpin 2012; Royston and others 2004]. Using a single life course variable (employment), Halpin [Halpin 2012] illustrated the benefit of multiple imputation of missing values in minimizing biases in clustering. However, he only studied gaps of multiple states occurring in the middle of the sequences and not at the beginning or the end. Also, despite of no previous evaluation, imputation of short survey gaps on single-variable sequences can be easily extended to multiple sequences.

For *alignment missing* and long internal data gaps that consist of more contiguous missing values, the amount of information in the sequence is often not enough to impute the gaps. The usual treatments were to include the missing values as a separate state [Aisenbrey and Fasang 2010] and/or apply normalization [Elzinga 2014]. These practices were criticized being problematic [Studer et al. 2018] as the clustering results were often based on the length of the sequences (i.e., observation time) rather than observed values. To mitigate such biases, most of the studies analyze only complete trajectories to preserve the holistic perspective. However, such practice will cause the sample size to be reduced and/or age cohorts and age spans biased towards those with all trajectories being fully observed [Studer et al. 2018].

According to review by Aisenbrey and Fasang [Aisenbrey and Fasang 2010], the sequence analysis literature is sparse when it comes to investigating to what extend such unavoidable contiguous gaps, especially in the form of *alignment missing* due to un-equal sequences, will change or bias clustering results and how such bias can be mitigated. The distance/dissimilarity measures calculated based on the OM procedure is deeply affected by the costs assigned to the indel and substitution transformations involving missing value states. Stovel and Bolan [Stovel and Bolan 2004] suggested lowering the indel only when dealing with incomplete sequences in order to reduce the biases caused by un-equal sequence lengths. However, tuning missing value substitution costs, which is for the first time proposed and utilized by this paper, provides the flexibility of adjusting distance contribution from missing values without affecting the transformation (indel or substitution) taken for non-missing values.

The OM distance structure of the data, in the form of a distance/dissimilarity matrix, that determine the clustering is critical for understanding the root causes induced by missing values. However, such distance structures are challenging to visualize due to high-dimensionality of the time series data. Various dimension reduction techniques have been proposed including linear and nonlinear mapping to the lower dimension spaces. Linear techniques that focus on preserving global data structures, such as Principal Component Analysis (PCA) [Hotelling 1933] and classical multidimensional scaling [Torgerson 1952], provide a good representation for linear data but do not work well when the underlying structure of the data is more complicated. For such data, nonlinear techniques that aim to preserve local data structures, are more appropriate. Nonlinear manifold methods include methods such as non-metric MDS [Kruskal 1964a,b], kernel PCA and Spectral Embedding [Belkin and Niyogi 2002; Schőlkopf et al. 1997], and t-distributed Stochastic Neighborhood Embedding (t-SNE) [Maaten and Hinton 2008]. Piccarreta and Lior [Piccarreta and Lior 2010] employed classic MDS for exploratory analysis of sequences. So far none of the nonlinear techniques have been applied to social sequence analysis to demonstrate their potential usage, which will be explored in this paper.

3 DATA DESCRIPTION AND PREPROCESSING

We analyze real world data extracted from the Panel Study of Income Dynamics (PSID) [PSID 2017] which includes a rich set of demographic and socio-economic information repeatedly collected from a large population of individuals tracked over multiple years, from 1968 through 2015. The initial dataset contained over 17,000 records for individual people, with missing values both over time and within each variable (i.e. some individuals only responded to the survey in some years, some individuals were not presented with certain survey questions, and individuals could choose not to answer any of the questions). Several preprocessing steps were needed to clean and transform the downloaded data to the format required for analysis.

Age is one of the variables collected every year the survey was conducted, which is used to align the sequences. The age variable was prone to noise and missing data, given human error and the timing of the survey relative to a respondent's birth month within a given calendar year. Using the age values collected over time, we calculated the birth year for each individual and used that to make subsequent corrections on the age variable.

For the joint sequence analysis, we considered a combination of five variables: two nominal (*family size* and *number of children under 8*), and three binary variables (*employment status*, *high school degree* and *marriage status*). The two nominal variables were represented by seven different states or values, including missing values. Life courses of these five variables are constructed by aligning the sequences by age between 20 and 60 for each individual in the final dataset. We aligned the data by age instead of calendar year in order to identify patterns in life course trajectories over individual lifetimes independent of the calendar years in which the relevant lifecycle events occurred.

After cleaning, we selected 1034 individuals whose data contained more than 23 contiguous non-missing values to be the focus of this study. This procedure limits survey gaps to short imputable ones for subsequent analysis. It is worth noting that, due to the necessity of identifying individuals with complete sequences between ages 20 and 60, and without excessive numbers of missing values, the resulting subsample is highly selected and not likely representative of the population or the sample of respondents to the PSID overall. The primary focus of this work is to compare methodologies for categorizing patterns in this structure of data. We therefore are not focusing on the representativeness of those patterns in the overall population at this point.

Because we aligned the sequences by age and focus on the age range between 20 and 60, individuals who are older than 20 at the beginning of the survey collection effort (year 1968) miss the leading part of the sequences (*alignment missing* or left censoring). Short survey gaps are present at the ending part of the sequences because after 1997 the survey was conducted only every two years rather than annually.

Figure 1 provides a plot of the family size variable for the final sample of 1034 individuals used in the analysis. It illustrates the mixed types of missing data arising in these data sources: short survey gaps (one value missing), and contiguous missing due to alignment by age. We also provide sample sequences for each of the five variables used here in the Supplementary Figure B.1. The missing patterns are similar for four of the variables: family size, number of children under 8, high school degree and marriage status. Only the employment variable has large chunks of missing values on both sides.

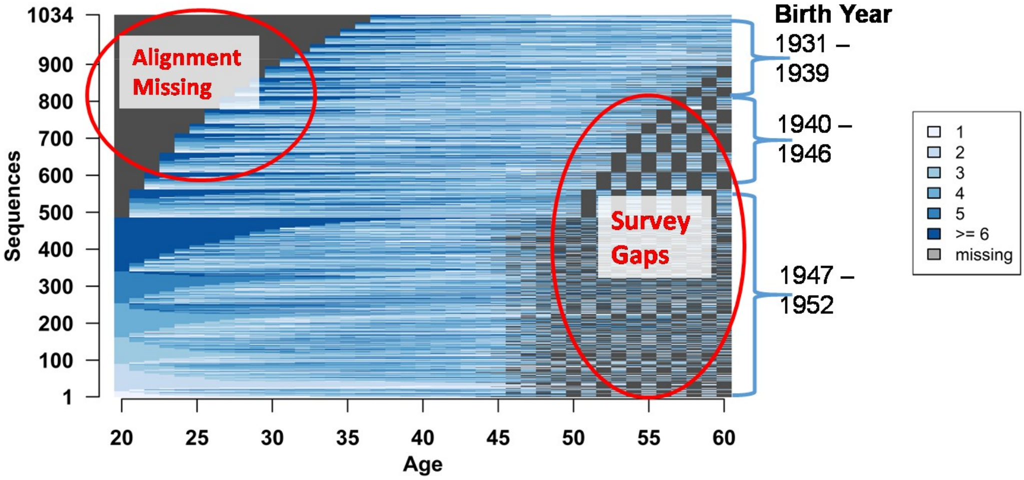

Fig. 1. Family size sequence representation for all the individuals in the dataset, to illustrate the missing value patterns that arise from survey gaps and missing segments after alignment by age.

4 METHOD

4.1 Distance Measures for Clustering

The OM distances described in the Related Work section were initially designed for one-variable categorical sequences. To extend this idea to multivariate sequences that include, for example, employment status, education, marriage and number of children, a new dissimilarity matrix has to take into consideration the contribution of each included variable. For this procedure, the indel cost is set to 1 and substitution costs are determined and set independently for each individual variable. We employ a data-driven approach by assigning substitution costs according to the transitional frequency between given states [Piccarreta and Billari 2007; Stovel et al. 1996]. Next, the substitution costs for the multivariate distance is calculated by averaging the substitution costs for the individual variables. This joint analysis approach follows [Pollock 2007] and is performed using the R package TraMineR version 2.0-6 [Gabadinho et al. 2011]. The choice for the substitution cost of the missing values (referred to as *NA cost* hereafter) is further explained in the next section.

4.2 Clustering Experiments

To systematically assess the effects of missing values and variable types on clustering solutions in the context of distances derived from Optimal Matching, we design 12 experiments.

As mentioned earlier, there are two types of missing values in our dataset, the treatment of which are described below.

Short *survey gaps*: These gaps can be addressed by imputation. The best predictors for imputing missing values are those observations that are the closest on the timeline to those that are missing, therefore for these one value internal gaps, our approach was to fill the gap with the value of the preceding immediately adjacent value (the *No Survey Gaps* case). Alternatively, missing can be included as a special state in OM with a user defined substitution cost for missing values (*Survey Gaps* case).

Alignment missing: these contiguous missing values are observed at the beginning of the sequences and not enough information is available for imputation. The treatment of *alignment missing* therefore consists of including ”missing” as a special state with a user defined substitution cost for missing values. The *alignment missing*, if applicable, is always present and therefore a substitution cost for the missing values (i.e. *NA cost*) must inevitably be chosen.

The default *NA cost* is set to 2 in OM [Gabadinho et al. 2011]. In this case, the cost of treating alignment missing is maximized. In contrast, in separate experiments, we set the NA cost to 0, which eliminates the cost of transforming any paired segments involving missing values from one to another. In this case, the cost of treating alignment missing or any survey gaps is minimized. In light of the above reasoning, we have 4 cases regarding treatment of missing values:

$$\{Survey\ Gaps,\ No\ Survey\ Gaps\} \otimes \{NA\ cost = 2, NA\ cost = 0\}$$

To construct complete experiments, we apply these 4 missing value treatment cases to three data domains of distinct joint sequence data types: (1) binary domain (employment, marriage, education); (2) nominal domain (number of children under 8, and family size); and (3) both of the above combined. The full 12 combinatorial set allows for systematic comparison of the effects of both missing values and mixed data types in relation to each other on life course trajectory clustering. Note here, missing values are in practice handled with same strategy across variables within the joint domain. Therefore, we do not consider further varying missing value treatments within the same data domain mentioned above.

4.3 Comparison Metrics

4.3.1 Cluster Quality Metrics. For clustering long-term life-course sequences, usually there is no ”ground truth” to be used for the direct evaluation of the proposed method. In this case, to evaluate clustering algorithms, several internal measures have been proposed to provide a statistical quality measure for the generated partitions, two of which are explained in detail below and utilized in our analysis. Internal clustering measures [Studer 2013] not only evaluate the quality of the returned clustering structure with no external help, but can be used to choose the best clustering algorithm and the optimal number of clusters for a given problem.

Hennig and Liao [Hennig and Liao 2010] suggest using Pearson’s correlation to evaluate and compare cluster solutions, which is an internal measure also known as ”Point Biserial Correlation” (equation 1). PBC is an index that is an easy measure of the resemblance between the distance matrix and the resulting hierarchical clustering dendrogram. This index measures the correlation of the distance matrix d with a matrix consisting of zeros and ones indicating whether two objects are in the same cluster or not and represented by a binary matrix d_{bin}. Let s_d and $s_{d_{bin}}$ be the standard deviation of d and d_{bin} respectively, and $s_{d,d_{bin}}$ be the covariance between d and d_{bin}. Given the above notations the PBC is computed as follows:

$$PBC = \frac{s_{d,d_{bin}}}{s_{d_{bin}} \cdot s_{d_{bin}}} \tag{1}$$

The Average Silhouette Width (equation 2) validates clustering performance based on the pair-wise difference of between- and within-cluster distances. Originally proposed by Kaufman and Rousseeuw [Kaufman and Rousseeuw 1990], this index is based on a notion of coherence of the assignment of an observation to a given cluster. This coherence is measured by comparing the average distance of an observation to the other members of its group with the average weighted distance to the closest group.

$$ASW = \frac{1}{NC}\sum_{i}(\frac{1}{n_i}\sum_{x \in C_i}\frac{b(x)-a(x)}{max(b(x),a(x))}) \tag{2}$$

$$a(x) = \frac{1}{n_i - 1}\sum_{y \in C_i, y \neq x} d(x,y) \tag{3}$$

$$b(x) = min_{j,j \neq i}\frac{1}{n_j}\sum_{y \in C_i} d(x,y) \tag{4}$$

Where NC is the number of clusters, n_i, is the number of objects in cluster i, C_i denotes cluster i, and $d(x, y)$ is the distance between x and y.

Additional measures are available but these two measures provide an objective way to choose the best combination of both the clustering algorithm and the number of clusters. Once the optimal number of clusters is selected, these parameters are used to generate the clustering groups.

4.3.2 Mutual Information Between Two Clustering Solutions of Different Missing Value Treatments. Comparison of the clustering results between any two experiments can be done by visual inspection and examination of membership distribution changes across the cluster solutions through cross tabulation as was previously done in [Halpin 2012]. Given the large number of experiments we intend to compare, we employ a simple metric, called normalized mutual information, to quantitatively assess how much the clustering solution changes from one treatment to another. Mutual information between two clustering solutions (R and L) can be computed from their contingency table by interpreting it as a table of joint probabilities $p(R, L)$. The probability of each cluster label can be computed by Equation(5) and Equation(6).

$$p(L) = \sum_{L} p(R,L) \tag{5}$$

$$p(R) = \sum_{R} p(R,L) \tag{6}$$

From these probabilities we compute entropies $H(R)$ 7 and $H(L)$ 8 and their mutual information $MI(R;L)$ 10.

$$H(R) = -\sum_{R} p(R) \cdot log((p(R)) \tag{7}$$

$$H(L) = -\sum_{L} p(L) \cdot log((p(L)) \tag{8}$$

$$H(R;L) = -\sum_{R}\sum_{L} p(R,L) \cdot log((p(R,L)) \tag{9}$$

$$MI(R;L) = H(R) + H(L) - H(R;L) \tag{10}$$

Finally, for ease of cross comparison, we use normalized mutual information (nMI) defined in (11).

$$nMI(R;L) = \frac{MI(R,L)}{\frac{1}{2}[H(R) + H(L)]} \tag{11}$$

4.4 Dimension Reduction Techniques

As the input data are 1034 individuals each associated with five categorical sequences, visualizing and diagnosing the clustering results of all 12 distance experiments in the original data space is challenging. In order to facilitate assessing clustering performance, cluster membership assignment, and their associations in different data domains, we employ dimension reduction techniques to display data points on a two-dimensional space. Dimensionality reduction is a process used to translate data from a high dimensional space to a low dimensional space with the goal of selecting the most important variables, and extracting relevant information or visualizing data in a meaningful way. Given the complexity of our dataset we considered only approaches to dimensionality reduction that are able to derive meaningful non-linear representations. Among non-linear algorithms, manifold learning methods, such as Multidimensional Scaling (MDS), Spectral Embedding and t-SNE have recently attracted great attention by providing excellent results on artificial and real-world datasets. [Bengio et al. 2006] provides an excellent review of them. We apply all of these dimensionality reduction techniques and select the most suitable one to enable us to assess and visualize the results of the clustering experiments.

4.4.1 Kernel Principal Component Analysis (kernel PCA) and Spectral Embedding (SE). Kernel PCA [Schőlkopf et al. 1999, 1997] is a non-linear dimensionality reduction technique well-known for its ability to deal with complex data. This method is an extension of the linear dimensionality reduction technique called Principal Component Analysis, which works by transforming the high dimension data into a lower dimension linear subspace using eigenvectors and eigenvalues. Kernel PCA works in two steps, first it redefines the input space using the kernel function k and second by performing PCA in the new feature space.

This idea was first applied as part as of the popular classification algorithms known as support vector machines (SVM). Kernel PCA uses a function Φ to transform the input vector x as a new vector $\Phi(x)$ whose size is n by n, where n is the number of input vectors from the dataset. Depending on the choice of function Φ, the data transformed in this higher space may become linear allowing us to perform the standard PCA to obtain the set of reduced feature vectors,χ. In the kernel space there is no need to compute the covariance of the Φ vectors, but we can compute their inner products of these vectors instead. Again, thanks to the kernel ”magic”, only the dot product of the transformed vectors is needed during the PCA step. Kernel PCA provides good results especially for non-linear data that lies along a manifold, however if the dataset includes too many data points it becomes computationally expensive to apply PCA to the kernel matrix.

Spectral Embedding (SE) [Belkin and Niyogi 2002], also known as Laplacian eigenmaps, is another approach to calculate non-linear embeddings that only preserves local distances. It first builds a similarity graph Laplacian to capture the local, neighborhood information existing in the dataset. This graph can be considered as a discrete approximation of the low dimensional manifold in the high dimensional space. As the second step it solves a generalized eigenvalue problem for the Laplacian matrix. The optimization process applied on the graph guarantees that the points connected on the graph Laplacian are mapped close to each other in the low dimensional space, preserving the local distances.

4.4.2 Non-metric MDS. The multidimensional scaling (MDS) method [Kruskal 1964a] is considered a classical statistical technique by now and it is usually employed for analyzing and visualizing the similarity or dissimilarity of data in a geometric space. The non-metric MDS version [Kruskal 1964b] identifies a non-parametric monotonic relationship between the data points using the affinity matrix. The algorithm is based on isotonic regression, which minimizes the stress function. In the stress formula 12, x represents the vector of similarities, $f(x)$ denotes a monotonic function and d is the matrix of distances between the data points.

$$Stress = \sqrt{\frac{\sum(f(x) - d)^2}{\sum d^2}} \tag{12}$$

The goal of the algorithms is to preserve the initial distances from the high dimensional space in the low dimensional space as much as possible. First the optimal monotonic transformation of the proximities has to be found. Secondly, the points of a configuration have to be optimally arranged, so that their distances match the scaled proximities as closely as possible. Additionally, MDS can also be computed even if the data matrix is unknown, all is needed is the Gram matrix.

4.4.3 t-SNE. t-SNE is one of the newest dimensionality reduction methods [Maaten and Hinton 2008], known for its good visualization capabilities. This method preserves the significant overall topology of the data points in the high-dimensional space when transformed to the low dimensional space. The algorithm works in two steps: first, it converts the similarities between data points to conditional probabilities; second, the stochastic neighbor embedding minimizes the sum of Kullback-Leibler divergences between the conditional probabilities of the low-dimensional embedding and the high-dimensional data using a gradient descent method. The t-SNE cost function 13 is not convex, therefore different initializations provide different output results.

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \tag{13}$$

In formula 13, P_i are the conditional probability distributions over all other data points given data point, and Q_i are the conditional probability distributions over all other mapped points given the map point in the lower dimensional space. t-SNE is often very successful at revealing clusters and subclusters in data, however it is not very good in at preserving the real distances between the data points that are far apart.

5 RESULTS AND DISCUSSIONS

5.1 Number of Clusters

Following previous research by [Piccarreta 2017] and [Pollock 2007] all the experiments were run using the Ward 's linkage hierarchical algorithm applied to multichannel distance matrices.

First, we report the Point Biserial Correlation (PBC) and the Average Silhouette Width Index (ASW) cluster validity indexes for the number of clusters k taking values in the [1, 10] interval. One multichannel distance matrix was computed for each of the 12 combination cases described in section 4.2. These distance matrices were then used as inputs for the clustering evaluation procedure. The results are presented on Figure 2.

For both PBC and ASW indexes the optimal clustering number is determined by maximizing the value of the index. Both index plots suggest that the best results are obtained for the distance matrices determined in the binary data domain (D_{binary}), followed by the combined domain ($D_{combined}$) and ending with the nominal data domain ($D_{nominal}$).

The PBC index clearly shows that for the nominal domain better clustering results are obtained when the cost of NA is set to 2. PBC indicates that in the combined and nominal domains it is harder to cluster when the NA cost is 0. Higher values for the ASW index are obtained in the binary domain when the NA cost is set to 0. The curves for the combined and nominal domains are overlapping in terms of Gaps and NA cost differences.

The best number of clusters indicated by the PBC and ASW measures varies not only with the domain, but it is also affected by the missing gaps and the NA cost choice. The plots suggest that most of the curves stabilize when k takes values between 3 and 5, and other than the nominal curves for NA cost 0, all of the curves start to decrease when k is greater than 5. These results have motivated our choice to run the further experiments using 4 clusters.

5.2 Choice of Dimensionality Reduction for Visualization

The clustering solutions can be examined in the original data space for each of the 12 experiments. An example is provided in Figure 3 using the nominal domain under ”with Gap, NA cost = 0” condition. Sequences of two nominal variables: family size (”Total in FU”) and number of children (”Children under 8”) are plotted separately for each of the clusters. The cluster solutions are visualized by (1) the state distribution by age in Figure 3 (a) and (2) the individual sequences in Figure 3 (b). Example visualization of the combined domain can be found in the Supplementary Figure B.2. This type of visualization is useful for interpreting the clustering results. For example, we can see that cluster 1 is dominated by individuals with large families having kids early in life, while individuals in cluster 2 individuals tend to have medium-sized families and have kids later and so on. However, for examining all 12 experiments and diagnosing their inter-relationship among various missing value and data type conditions, these visual tools are not very effective.

We explore 4 different types of dimension reduction techniques with the goal of finding the most effective one for visualizing the internal data structure represented in the distance matrices as captured by the clustering solutions. Dimension reduction techniques map the data points onto a two-dimensional space for easy visualizations. We color the points by clustering labels derived for each of the 12 experiments. Both Kernel PCA and SE fail to capture any clustering patterns in the lower dimensions. t-SNE appears to best capture the ”closeness” pattern of the points assigned to the same clusters (Figure 4) across all the data domains. MDS also performs well in the binary and combined domains but fails to capture the high dimensional data structure in the nominal domain. From the reduced dimension plots of t-SNE and MDS we can see more dispersed patterns within the nominal domain relative to binary and combined domains, which explains lower performance metrics of clustering in Figure 2.

The t-SNE approach focuses on capturing the local structures of the data. The resulting topology and the global geometry (e.g. distances between clusters) are sensitive to the hyperparameters used. We examine an array of hyperparameter combinations and determine the ones that have the most stable performance and can best capture both local and global geometry as represented by the clustering results. Specifically, t-SNE mapping of original distance matrices to two-dimensional space is sensitive to two input hyperparameters: maximum number of iterations and perplexity number. The perplexity number is usually interpreted as a smooth measure of neighborhood size around a given data point and it should be smaller than the average size of the clusters. Given the size of our clusters are usually greater than 150, we vary the perplexity from 10 to 110. We also examine how the t-SNE results change with the maximum number of iterations by varying it from 500 to 1000 (Supplementary Figure B.3). We find that smaller sub-clusters appear within the same cluster when perplexity is small (10 or 30), while such pattern disappears when perplexity

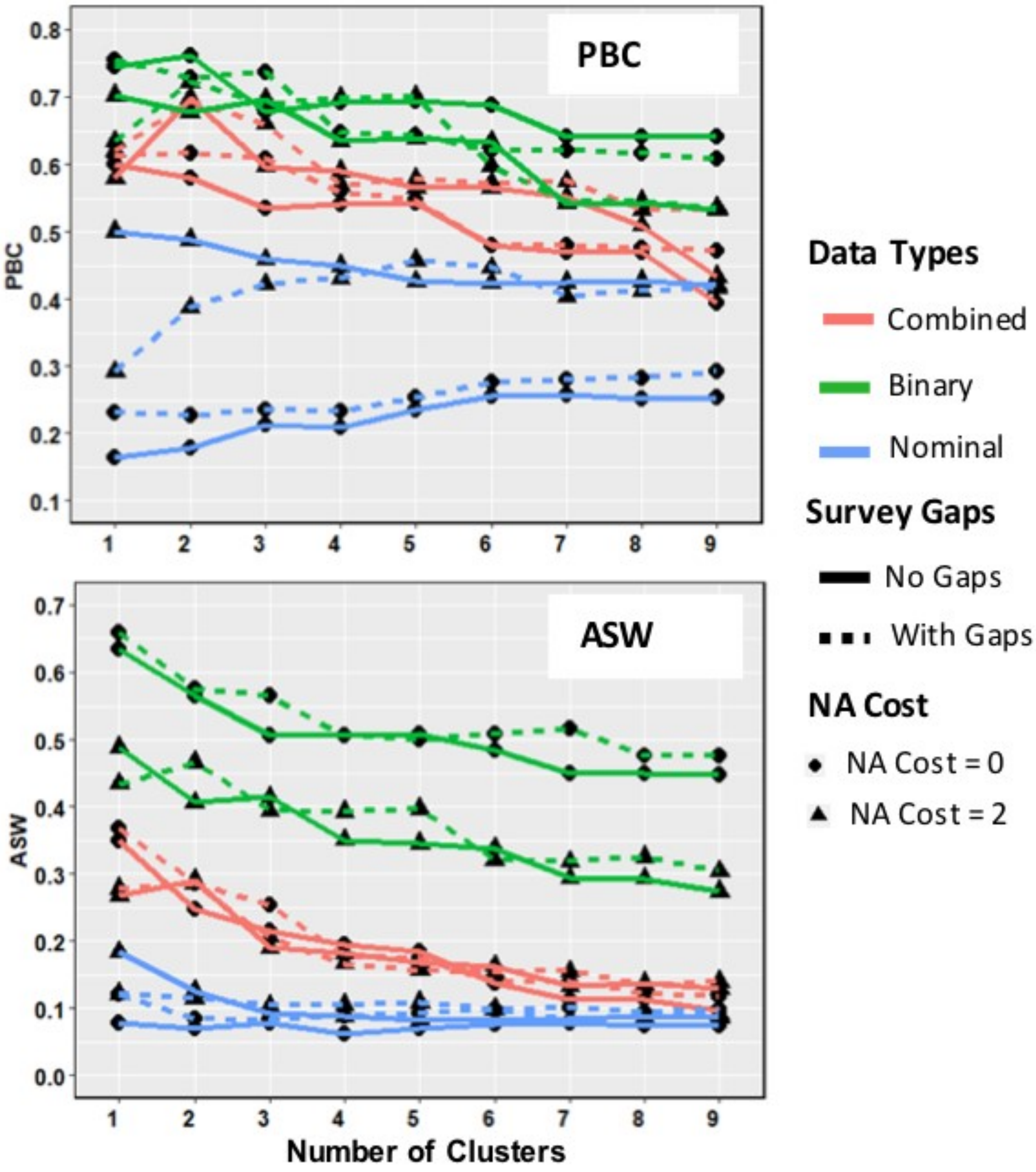

Fig. 2. Cluster validity metrics PBC and AWS as a function of number of clusters.

numbers are greater. This suggests that the choice of perplexity number should be greater than 30. There is also evidence that iteration = 500 can sometimes produce outlier patterns as shown in perplexity = 30 and 50 cases under ”with gap and NA cost = 0” conditions for the combined domain.

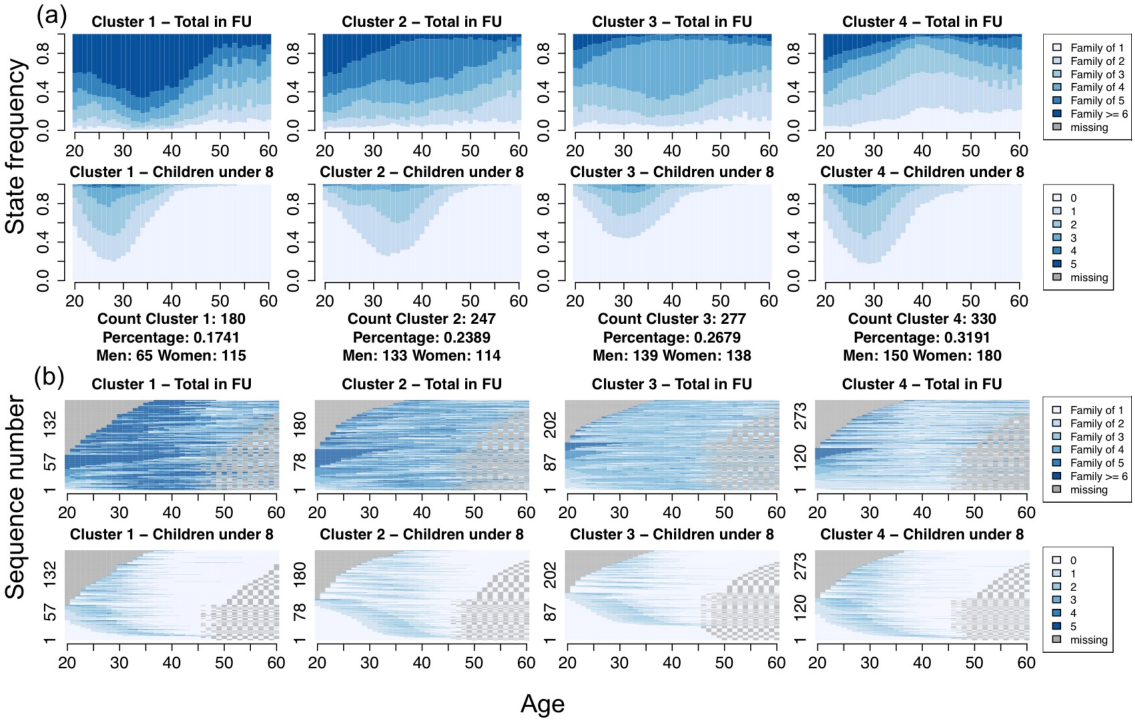

Fig. 3. Cluster solutions in the $D_{nominal}$ with survey gaps and $NAcost = 0$. (a) sequence state distribution by age; and (b) individual sequence by age. Variables shown are: family size ("Total in FU") and number of children ("Children under 8").

In general, perplexity choice of 90 produces the most stable and compelling representation of the cluster patterns in that points of the same cluster labels are generally close together. Therefore, we choose maximum iteration number of 1000 and perplexity number of 90 for subsequent analyses.

5.3 Effects of Missing Values on Clustering

In this section, we examine the effects of missing value treatments on clustering solutions in relation to variable type selection (binary, nominal, and combined) and their respective resulting dissimilarity structures within the data. In general, we find the clustering results derived from the nominal domain are most sensitive to missing value treatments especially the choices of NA cost. More importantly, aided by dimensionality reduction, we are able to identify artifacts in dissimilarity structures and associated clustering solutions driven by age cohorts when NA cost is maximized. More detailed comparison is presented below.

”Gap vs No Gap” (first two rows in Table 1) represents the commonality (measured by *nMI*) observed in clustering solutions between cases where short survey gaps are imputed and not imputed. Lower values of *nMI* indicate greater differences and thus greater effects of the gap imputation treatment. As discussed earlier, changes due to short gap imputation need to be evaluated conditioned on the choice of NA cost because alignment missing is present in all cases.

The ”Gap vs No Gap | NA Cost=2” case measures the effect of gap imputation when the influence of alignment missing on distance measures are maximized, while the ”Gap vs No Gap | NA Cost=0”

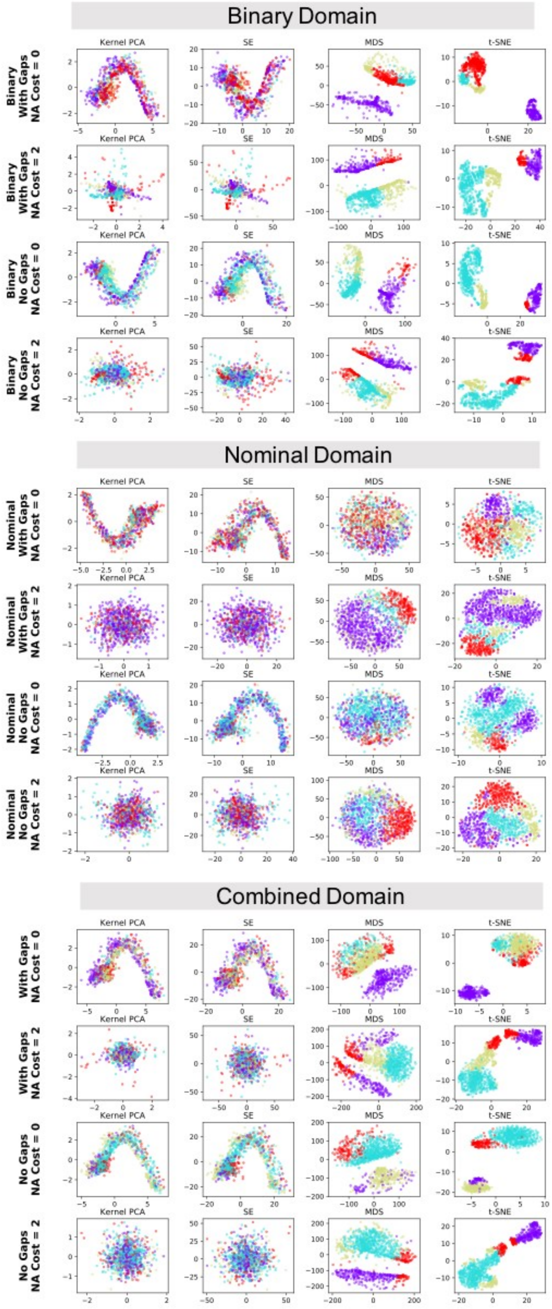

Fig. 4. Clustering results visualized by dimensionality reduction techniques. Data points are color coded by cluster memberships within respective data domains (binary, nominal, and combined). Dimension reduction techniques applied from left to right: Kernel PCA, Spectral Embedding (SE), Non-metric MDS, and t-SNE.

case measures the same effect when the influence of alignment missing is minimized. In general, we see imputation of short survey gaps changes clustering in $D_{nominal}$ much more than in D_{binary} or in $D_{combined}$. Imputation effects in the $D_{nominal}$ are also sensitive to the choice of NA substitution cost. In $D_{nominal}$, clustering solutions using data with and without survey gaps become more different when NA cost changes from 2 to 0.

”NA cost = 2 vs NA cost = 0” (bottom two rows in Table 1) represents the commonality between the choices of NA cost specification when missing values are included as a special category. In Table 1, the ”NA cost = 2 vs NA cost = 0 | no gap” case represent NA cost effects due to alignment missing alone, while ”NA cost = 2 vs NA cost = 0 | gap” case represents NA cost effects due to the presence of both short survey gaps and alignment missing. Similar to the imputation effects, the choice of NA cost affects clustering solutions derived from $D_{nominal}$ the most, especially when survey gaps are also present. $D_{combined}$ also becomes more sensitive to the choice of NA cost when survey gaps are present. Therefore, imputation of short survey gaps helps stabilize clustering solutions.

Table 1. Normalized Mutual Information *nMI* Between Clustering Solutions Derived With Different Missing Data Treatments (<0.5 values are masked in pink)

Comparison of Treatments	***Data Domain***		
	D_{binary}	$D_{nominal}$	$D_{combined}$
Gap vs No Gap | NA Cost=2	0.64	0.36	0.71
Gap vs No Gap | NA Cost=0	0.62	0.13	0.63
NA cost = 2 vs NA cost = 0 | No Gap	0.66	0.26	0.50
NA cost = 2 vs NA cost = 0 | Gap	0.51	0.10	0.37

Due to alignment by age, as shown in Figure 1, the degree of missing data due to alignment missing and/or survey gaps are largely driven by individual’s birth year timing. Age cohorts born 1947-1952, 1940-1946, and 1931-1939, respectively are subject to varying degrees of both alignment missing and survey gaps (Figure 1). To explore the potential biases in distance measures caused by the choice of NA cost, we examine these age cohorts within the data structures captured by the reduced dimension via t-SNE representation in Figure 5 (a) in $D_{combined}$. In doing this, it is evident that data points of the same age cohorts are located relatively close to each other when NA cost is 2. This means when NA cost is 2, the distance structure within the data is largely driven by the birth year patterns. In contrast, this effect disappears when NA cost is 0. Therefore, the missing patterns in the sequences as differentiated by the age cohorts have a clear impact on the distance measures when NA cost is maximized.

The birth year distribution of clustering solutions under 4 cases of missing value treatments further illustrates this effect in Figure 5 (b) using $D_{combined}$ as an example. A birth-year driven clustering solution is especially evident in the NA cost = 2 case when survey gaps are also present in addition to alignment missing. In this case, Clusters 2, 3, and 4 are driven by age cohorts born 1947-1952, 1940-1946, and 1931-1939, respectively. Effects from both types of missing data are maximized when NA cost = 2, leading to the most significant biases by age cohorts in the distance calculation. Such bias due to ”NA cost = 2” is alleviated when survey gaps are imputed. However, we can still observe Cluster 4 being largely driven by the age cohort born 1931-1939, due to the most serious alignment missing alone in this age cohort. These biases can be confirmed by comparison to the ”NA cost = 0” cases, where age cohort effects completely disappear as ”NA cost = 0” minimizes the contribution from missing values to the distance computation. We also observe similar behavior for the binary and nominal domains (Supplementary Figure B.4). Therefore, ”NA cost = 2” maximizes

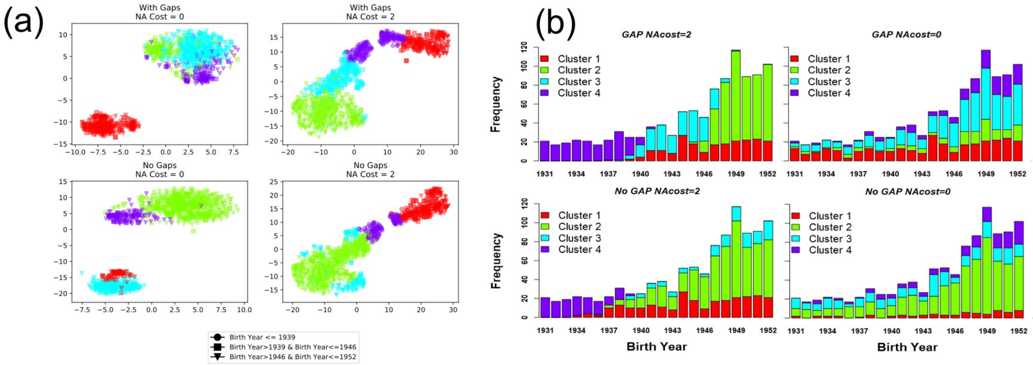

Fig. 5. (a) Birth year relations within the dissimilarity structure in $D_{combined}$. Colors indicate cluster membership and shapes indicate age cohorts; (b) Birth year distribution of cluster assignments derived from $D_{combined}$ under four missing value treatments: $\{Survey\ Gaps,\ No\ Survey\ Gaps\} \otimes \{NA\ cost = 2, NA\ cost = 0\}$.

the missing value contribution, meaning that cluster definition and assignment is being largely driven by age cohorts (i.e., alignment missing cohorts) rather than relevant information with the input data itself, leading to a prominent pattern shared by dissimilarity matrices of all domains.

5.4 Effects of Mixed Data Types on Clustering

The interpretation of the clusters in the combined domain can become problematic if they are not equally representative of all relevant contributing domains. Therefore, understanding how the data domains are associated with each other is critical for cluster interpretation. The best clustering solution of a specific domain ($D_{combined}$, D_{binary}, or $D_{nominal}$) represents an optimal simplification of respective dissimilarity structures. Therefore, domain associations in the presence of missing values can be investigated by: (1) the commonality among the best clustering solutions obtained from the combined and contributing domains, (2) the performance of clustering solutions derived from the combined domain on its contributing domains, and (3) how the previous two factors change with missing value treatments.

Figure 6 demonstrates that overall, the clustering solution derived from D_{binary} and $D_{combined}$ are more similar and they are both different from clustering derived from $D_{nominal}$, indicating greater association of the combined domain with the binary domain. We have seen in previous sections that distance matrices derived from the binary data domain $D_{combined}$ are easier to cluster than those from the nominal domain ($D_{nominal}$). The greater association of the combined domain with the binary domain observed here is consistent with joint sequence analysis literature that the clustering solution from a combined data domain may favor the contributing domains that are easier to cluster [Piccarreta 2017]. A more important observation is that the commonality of clustering solutions between $D_{nominal}$ and other domains is especially sensitive to the treatment choice of missing values (Figure 6). In general, commonality in clustering solutions between $D_{nominal}$ and other domains is greater under the ”NA Cost = 2” cases than the ”NA cost =0” cases (indicated by the darker blue first row of plots relative to the second row in Figure 6). In fact, the 4-cluster solution derived from $D_{combined}$ also shows better performance, as measured by PBC and ASW, on D_{binary} than on $D_{nominal}$, for all missing value treatment cases (Supplementary Table A.1)

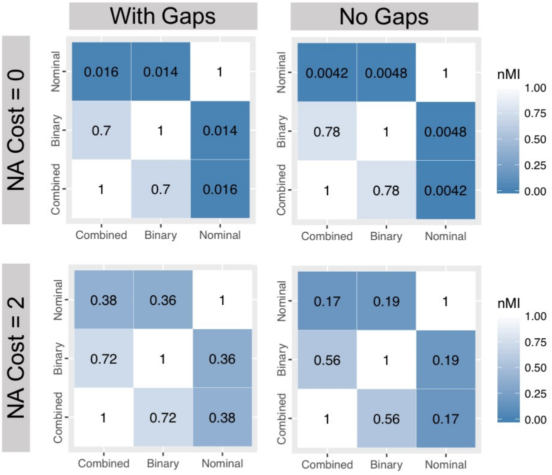

Fig. 6. Normalized Mutual Information (*nMI*) between clustering solution derived from binary, nominal, and combined domains under different missing value treatments. Darker blue indicates greater differences between the pair.

To visually diagnose domain associations, we color the data points in binary and nominal domains based on the cluster solutions derived from the combined domain for each of the missing value treatment experiment (Figure 7). The combined domain cluster solutions can largely capture the data structures in the binary domain as expected from the large *nMI* in the cluster solutions between the two domains. However, we can see that points of the same color (and therefore belong to the same cluster in the combined domain) are largely dispersed in the nominal domain when NAcost is 0. When NA cost is 2, the combined domain cluster solution starts to capture the clustered patterns in the nominal domain especially in the ”purple” area. This pattern underlies the reason why the association (indicated by *nMI*) between $D_{combined}$ and $D_{nominal}$ increases as NA cost changes from 0 to 2 in Figure 6. Birth year brushing in Supplementary Figure B.4 indicates that these purple points are largely ”older” people in the survey (birth year < 1939) and belong to the same cluster in the nominal domain.

When NA cost is 2, the cost to align segments involving missing values are maximized, which systematically increases the edit distance between sequences with and without certain missing segments. Consequently, sequences with similar missing patterns become relatively more similar. Such shared biases in distance measures lead to an artificial association between the domains and thus increases the commonality in clustering solutions. Joint sequence analysis is found mostly useful when the individual domains are interdependent or associated [Gauthier et al. 2010]. However, missing values can complicate these domain association patterns. This finding highlights the importance of the choice of missing value treatment when interpreting the clustering solutions in joint sequence analysis, as the association pattern can be entirely driven by missing values.

5.5 Mitigation of Alignment Missing

While imputation can mitigate short survey gaps, we have seen that alignment missing coupled with the choice of NA cost can significantly affect the dissimilarity measures and subsequently the clustering solutions (i.e. the age bias in clustering solutions). We have demonstrated that clustering in the nominal domain is especially sensitive to the choice of missing value treatment.

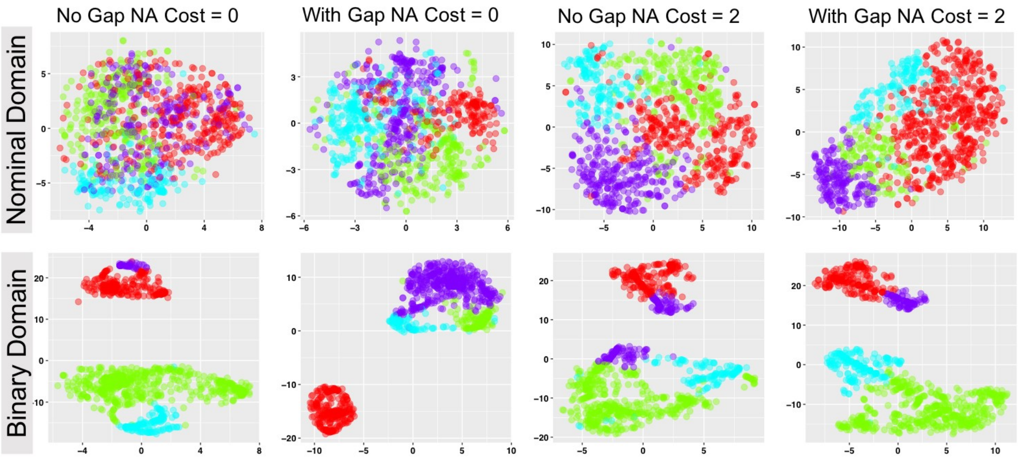

Fig. 7. t-SNE representation of data structures in the nominal and binary domains colored by the 4-cluster solution derived from the combined domain

The choice of NA cost we have considered so far are two extreme cases: (1) setting NA substitution cost to 2 is equivalent to considering missing values as a complete different state from other states and therefore requires the maximum ”edit” cost to transform any missing values to the target state values; (2) setting NA substitution cost to 0 is equivalent to treating missing values as a ”wild card” that is the same state to any given target state and therefore needs zero ”edit” cost to transform. We have shown that the first case introduces biases in dissimilarity measures. The second one is not very realistic in practice. In reality, the NA cost can be more flexible and take values between 0 and 2, effectively allowing some probability for missing values to be in one versus other of the non-missing states. The appropriate choice of NA cost can be identified on a case-by-case basis. From the perspective of eliminating biases caused by alignment missing, we can choose the proper NA cost value by varying it by small increments from 2 to 0 and identifying the one when the age biases start to disappear. In Figure 8 (a), we show that the missing value induced distance patterns (as colored by birth year) start to dissolve as NA cost decreases from 2 to 0.4.

Another data-driven way to mitigate the biases introduced by alignment missing is to find a more appropriate age span so that the consequence of contiguous missing patterns in the sequences no longer dominate the dissimilarity measures. Figure 8 (b) vary the age span by increasing the lower age limit from 20 to 35. We can see at lower age limit of 29 to 32, the artificial patterns induced by missing values begin to disappear.

These two data driven approaches graphically guided by t-SNE visualizations ensure the dissimilarity measures computed through optimal matching are no longer driven by missing values. The state distribution by age plots are shown in Figure 9 and individual sequence plots are shown in Supplementary Figure B.5. We can see the resulting trajectory patterns during the overlapping age span (32 to 60) are very similar between the two approaches, confirming that the patterns are mostly driven by the values in the input variables instead of the length of missing values.

Under the conditions of these two approaches, the combined domain shows little true association with the nominal domain. The *nMI* between the combined and nominal domains reduces from 0.17 (NA cost =2 and age span = 20 to 60) to 0.003 (NA cost =0.4 and age span = 20 to 60) and 0.008 (NA cost =2 and age span = 32 to 60). This means that the clustering solutions derived from the combined domain represent little patterns in its contributing nominal domain, which indicates that

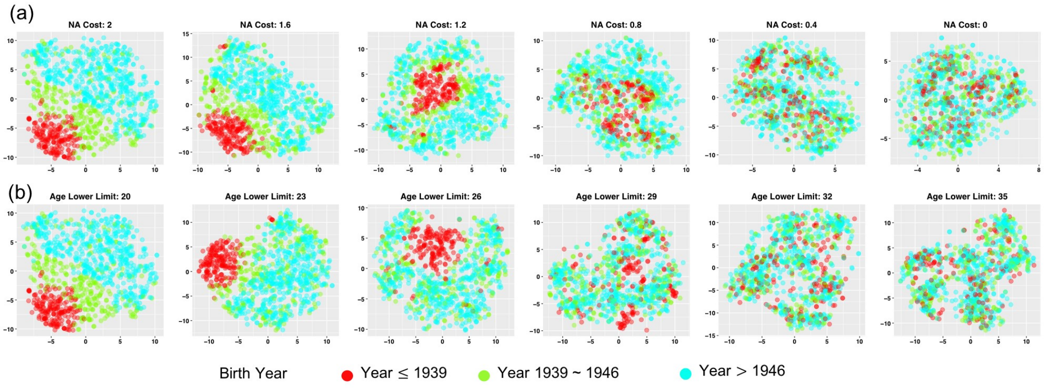

Fig. 8. Changes in data structures with varying (a) NA cost (from 0 to 2 at a 0.4 increment) and (b) age limits (starting age from 20 to 35 at a 3-year increment), visualized by t-SNE for the nominal domain after the short survey gaps are imputed. Data points are colored by age cohorts.

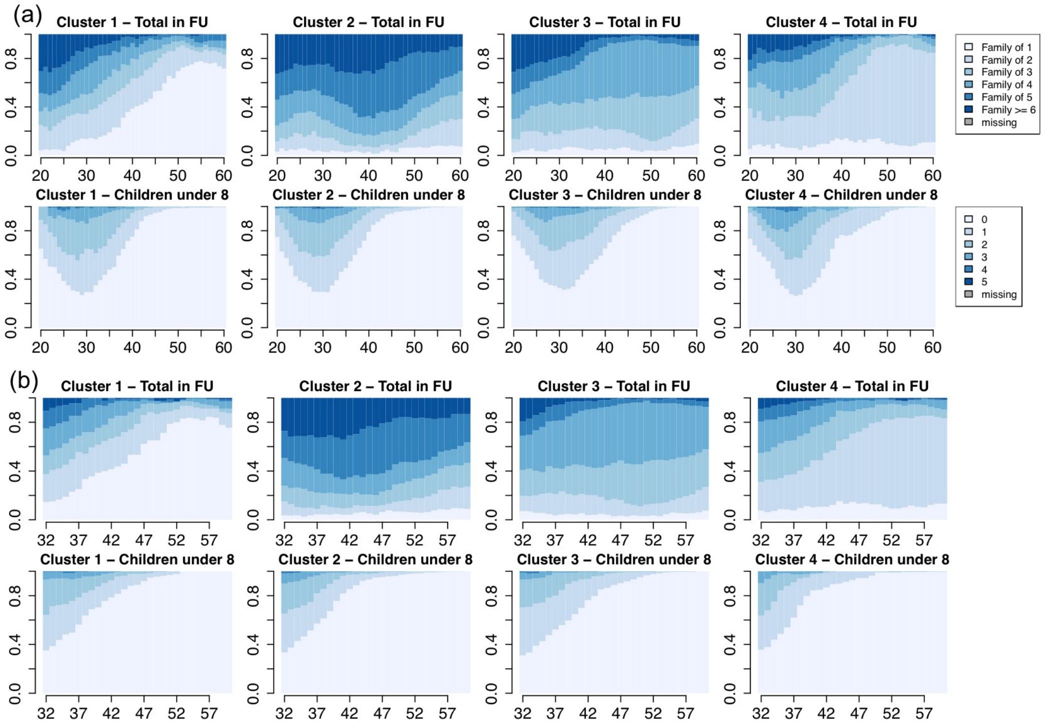

Fig. 9. State distribution plot of cluster solutions in the nominal domain under no survey gap condition for (a) NA cost = 0.4, age span = 20 to 60; and (b) NA cost = 2, age span = 32 to 60.

the binary and nominal variables used here should be ideally clustered separately to derive more meaningful and representative patterns.

6 CONCLUSIONS

This paper evaluates two issues facing joint social sequence analysis using real-world data: missing values and mixed data types. Changes in clustering solutions are systematically assessed in a full combination of experiments designed so that the two types of problems can be examined in relation to each other. We applied the experiments to a real world data set obtained from the PSID where both short and long data gaps are present due to either survey gaps or alignment by age. Past evaluation of missing values has focused on short gap imputation strategies in the context of single variable sequence clustering. Our study addresses the effects of ”unavoidable” missing values arising from sequence alignment in addition to imputable short gaps, and missing value effects are examined in relation to joint sequence analysis with various types of variables. Among the application of four nonlinear dimensionality reduction techniques, t-SNE is found most successfully visualize the clustering solutions. t-SNE is applied to visually display and diagnose the dissimilarity measures and their subsequent effects on clustering.

We find missing values and their choices of treatment (imputation and NA cost specification) mostly affects the clustering solution in the nominal sequences that have greater state spaces. Imputation of short survey gaps helps stabilize clustering solutions. With the alignment missing data problem, choices of NA cost are important. The traditional default way of including missing as a special state maximizes NA cost (=2), leads to artificial clusters driven by different cohorts based on the alignment dimension (in our case age), while setting NA cost to 0 eliminates such behavior. Maximizing NA cost in the presence of missing values is also found to inflate the ”apparent” cluster performance measured by quality metrics. Such treatment needs to be practiced with caution in future OM applications.

We find distance matrices derived from the binary data domain are easier to cluster than those from mixed or nominal data types. As a result, clustering solutions from the combined domain favors the contributing domain of binary variables, indicated by greater commonality in their respective optimal clustering solutions. However, association between the nominal domain and the combined domain can sometimes be artificially inflated in the presence of missing values together with a specification of high NA cost, leading to ”apparent” association between the two domains. This finding highlights the importance of the choice of missing value treatment in correct interpretation of the clustering solutions represented in the contributing domains.

We employ t-SNE dimensionality reduction to demonstrate two practical data driven approaches, tuning NA cost and identifying proper age span, to realistically mitigate serious biases in distance measures caused by missing values. Such approaches can at the same time maximize data utilization by avoiding removing entire sequences or age spans that involve missing values as commonly practiced in the current literature. The analysis framework illustrated here can be easily extended to different datasets and other variants of OM to guide proper missing value handling and result interpretation in social sequence analysis studies.

A SUPPLEMENTARY TABLE

Table A.1. Performance of the 4-cluster solution of $D_{combined}$ on its contributing domains (D_{binary} and $D_{nominal}$) under different missing value treatments

Contributing Domains	**With Gaps NACost=0**	**With Gaps NACost=2**	**No Gaps NACost=0**	**No Gaps NACost=2**
		Point Biserial Correlation		
D_{binary}	0.62	0.57	0.73	0.62
$D_{nominal}$	0.03	0.39	-0.05	0.21
		Average Silhouette Width		
D_{binary}	0.47	0.27	0.55	0.35
$D_{nominal}$	-0.03	0.07	-0.05	0.00

B SUPPLEMENTARY FIGURES

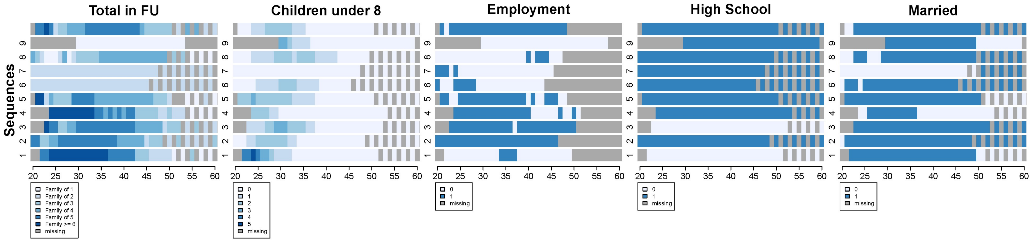

Fig. B.1. Sample individual sequence plots to illustrate the missing patterns in the PSID dataset. Variables are family size ("Total FU"), number of children under 8 ("Children under 8"), employment status ("Employment"), high school degree ("High School"), and marriage status ("Married").

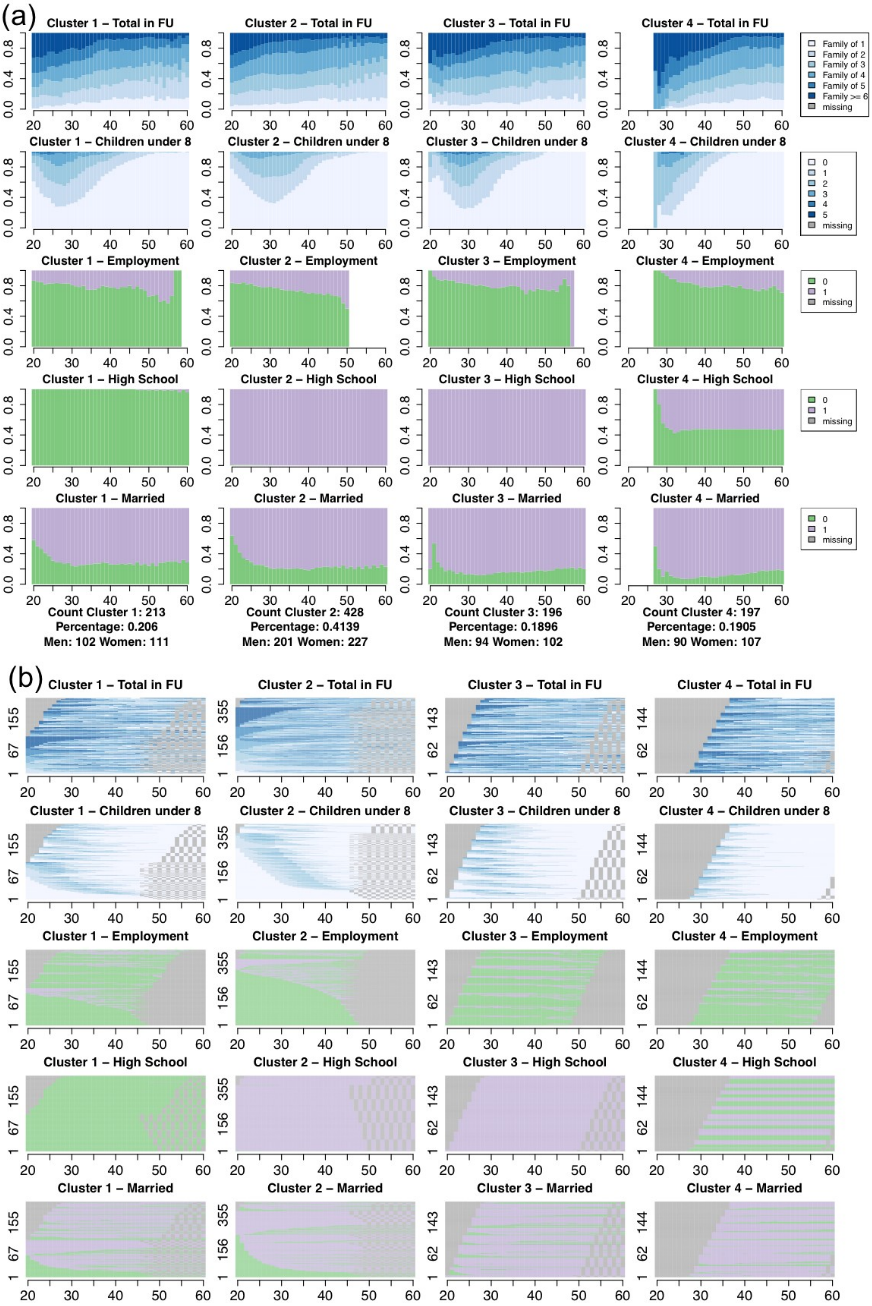

Fig. B.2. Cluster solutions for combined domain with survey gaps and NA cost = 2. (a) state distribution plot; (b) individual sequence plot.

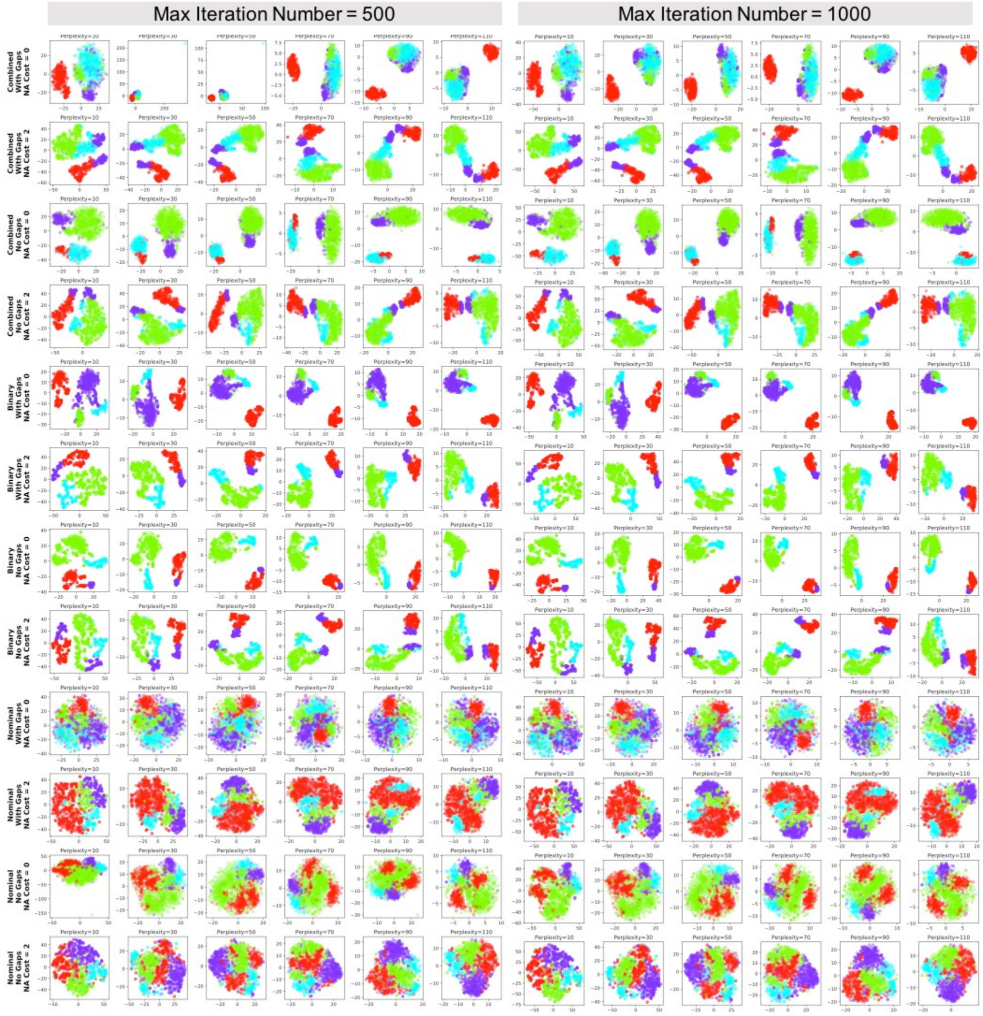

Fig. B.3. t-SNE results from varying perplexity number and maximum iteration number (colored by the clustering solutions).

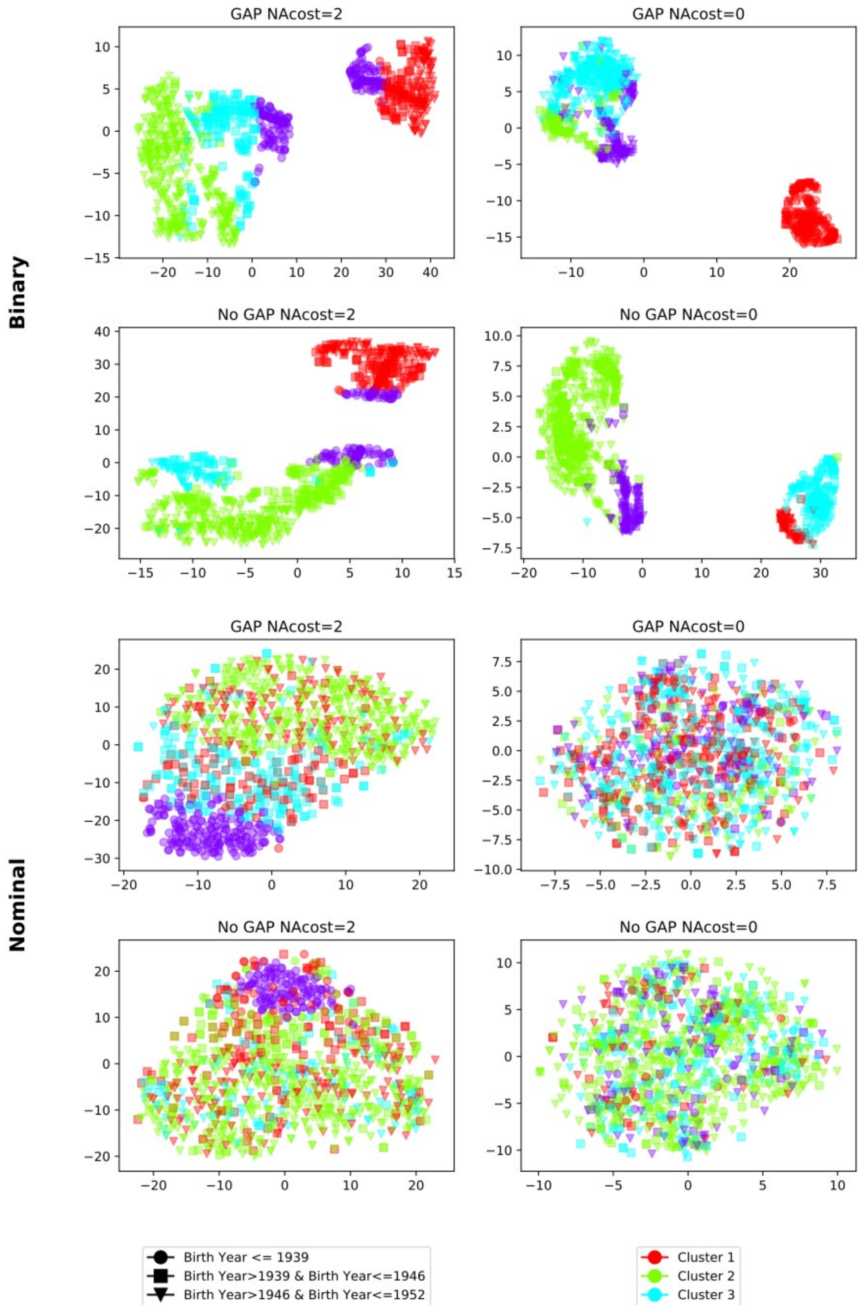

Fig. B.4. Birth year and cluster assignment relationship in binary and nominal domains.

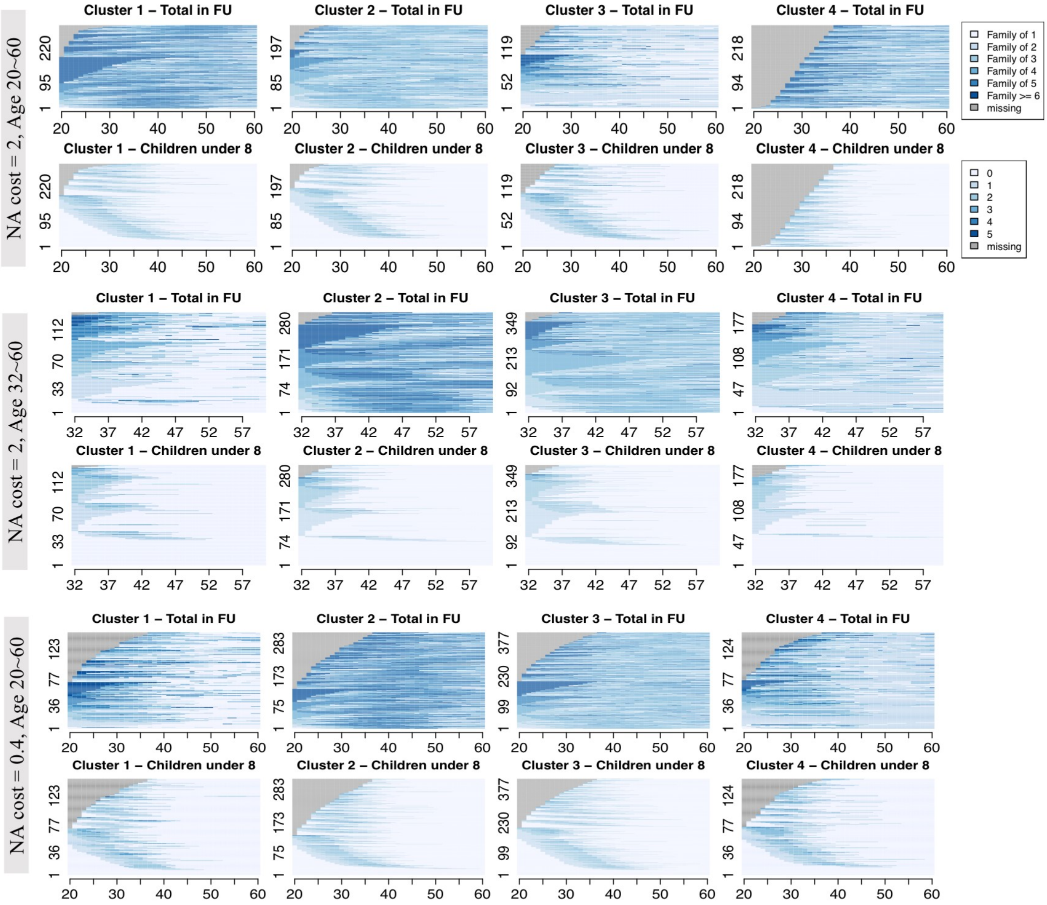

Fig. B.5. Individual sequence plot of cluster solutions for nominal domain under no survey gap cases.

ACKNOWLEDGMENTS

This paper and the work described were sponsored by the U.S. Department of Energy (DOE) Vehicle Technologies Office (VTO) under the Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility Laboratory Consortium, an initiative of the Energy Efficient Mobility Systems (EEMS) Program, and and Office of Science under Advanced Scientific Computing Research (ASCR) program, under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231. The authors acknowledge Anand Gopal of Lawrence Berkeley National Laboratory for leading the Mobility Decision Science Pillar of the SMART Mobility Laboratory Consortium. The following DOE Office of Energy Efficiency and Renewable Energy (EERE) managers played important roles in establishing the project concept, advancing implementation, and providing ongoing guidance: David Anderson, Rachael Nealer, and Jake Ward. This work was also supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Visiting Faculty Program (VFP).

REFERENCES

Andrew Abbott and John Forrest. 1986. Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History* 16, 3 (1986), 471–494. http://www.jstor.org/stable/204500

Andrew Abbott and Alexandra Hrycak. 1990. Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American journal of sociology* 96, 1 (1990), 144–185. http://www.journals.uchicago.edu/doi/abs/10.1086/229495

Silke Aisenbrey and Anette E. Fasang. 2010. New life for old ideas: The" second wave" of sequence analysis bringing the" course" back into the life course. *Sociological Methods & Research* 38, 3 (2010), 420–462. http://journals.sagepub.com/doi/abs/10.1177/0049124109357532

Mikhail Belkin and Partha Niyogi. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*. 585–591.

Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. 2006. Spectral dimensionality reduction. In *Feature Extraction*. Springer, 519–550.

Hilde Bras, Aart C. Liefbroer, and Cees H. Elzinga. 2010. Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography* 47, 4 (2010), 1013–1034. http://www.springerlink.com/index/3547442765W022X4.pdf

Cees H. Elzinga. 2007. *Sequence analysis: Metric representations of categorical time series*. Technical Report. Department of Social Science Research Methods, Vrije Universiteit, Amsterdam. https://www.researchgate.net/profile/Cees_Elzinga/publication/228982046_Sequence_analysis_Metric_representations_of_categorical_time_series/links/5464a15e0cf2c0c6aec64294.pdf

Cees H Elzinga. 2014. Distance, similarity and sequence comparison. In *Advances in sequence analysis: Theory, method, applications*. Springer, 51–73.

Tak-chung Fu. 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence* 24, 1 (2011), 164–181. http://www.sciencedirect.com/science/article/pii/S0952197610001727

Alexis Gabadinho, Gilbert Ritschard, Nicolas Séverin Mueller, and Matthias Studer. 2011. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software* 40, 4 (2011), 1–37. https://archive-ouverte.unige.ch/unige:16809

Jacques-Antoine Gauthier, Eric D. Widmer, Philipp Bucher, and Cédric Notredame. 2009. How much does it cost? Optimization of costs in sequence analysis of social science data. *Sociological Methods & Research* 38, 1 (2009), 197–231. http://journals.sagepub.com/doi/abs/10.1177/0049124109342065

Jacques-Antoine Gauthier, Eric D. Widmer, Philipp Bucher, and Cédric Notredame. 2010. Multichannel sequence analysis applied to social science data. *Sociological methodology* 40, 1 (2010), 1–38. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9531.2010.01227.x/full

Brendan Halpin. 2012. *Multiple Imputation for Life-Course Sequence Data*. Technical Report WP2012-01. Department of Sociology, University of Limerick.

Brendan Halpin. 2014. Three Narratives of Sequence Analysis. In *Advances in Sequence Analysis: Theory, Method, Applications*. Springer, Cham, 75–103. https://doi.org/10.1007/978-3-319-04969-4_5

Christian Hennig and Tim F. Liao. 2010. *Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification*. Technical Report 308. Department of Statistical Science, University College London. http://www.cnmd.ac.uk/statistics/research/pdfs/rr308.pdf

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24, 6 (1933), 417.

Ling Jin, Doris Lee, Alex Sim, Sam Borgeson, Kesheng Wu, C. Anna Spurlock, and Annika Todd. 2017. Comparison of Clustering Techniques for Residential Energy Behavior using Smart Meter Data. In *AI for Smart Grids and Buildings Workshop*. San Francisco, CA.

R. Burke Johnson and Anthony J. Onwuegbuzie. 2004. Mixed methods research: A research paradigm whose time has come. *Educational researcher* 33, 7 (2004), 14–26. http://journals.sagepub.com/doi/abs/10.3102/0013189X033007014

Leonard Kaufman and Peter J. Rousseeuw. 1990. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis* (1990), 68–125. http://onlinelibrary.wiley.com/doi/10.1002/9780470316801.ch2/summary

Dimitrios Kotsakos, Goce Trajcevski, Dimitrios Gunopulos, and Charu C Aggarwal. 2013. Time-Series Data Clustering. In *Data Clustering*. Chapman and Hall/CRC, 357–380.

Joseph B. Kruskal. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1 (1964), 1–27.

Joseph B. Kruskal. 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 2 (1964), 115–129.

Hugh Lauder, Phillip Brown, and A. H. Halsey. 2004. Sociology and political arithmetic: some principles of a new policy science. *The British Journal of Sociology* 55, 1 (2004), 3–22. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-4446.2004.00002.x/full

T. Warren Liao. 2005. Clustering of time series data — a survey. *Pattern recognition* 38, 11 (2005), 1857–1874. http://www.sciencedirect.com/science/article/pii/S0031320305001305

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48, 3 (1970), 443–453. http://www.sciencedirect.com/science/article/pii/0022283670900574

Raffaella Piccarreta. 2017. Joint Sequence Analysis: Association and Clustering. *Sociological Methods & Research* 46, 2 (2017), 252–287. http://journals.sagepub.com/doi/abs/10.1177/0049124115591013

Raffaella Piccarreta and Francesco C. Billari. 2007. Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170, 4 (2007), 1061–1078.

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

Gary Pollock. 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 (2007), 167–183. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2006.00450.x/full

PSID 2017. Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. (2017). https://psidonline.isr.umich.edu/

Sangeeta Rani and Geeta Sikka. 2012. Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications* 52, 15 (2012), 8887. http://search.proquest.com/openview/988426c32ceb2142080844b3487a5724/1?pq-origsite=gscholar&cbl=136216

Patrick Royston and others. 2004. Multiple imputation of missing values. *Stata journal* 4, 3 (2004), 227–41. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.466.8480&rep=rep1&type=pdf

Bernhard Schőlkopf, Christopher JC Burges, and Alexander J. Smola. 1999. *Advances in kernel methods: support vector learning*. MIT press.

Bernhard Schőlkopf, Alexander Smola, and Klaus-Robert Müller. 1997. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*. Springer, 583–588.

Reto Schumacher, Koenraad Matthijs, and Sarah Moreels. 2012. Migration and reproduction in an urbanizing context. a sequence analysis of family life courses in 19th century Antwerp and Geneva. (2012). https://lirias.kuleuven.be/bitstream/123456789/345904/1/WOG+working+paper+17.pdf

Katherine Stovel and Marc Bolan. 2004. Residential trajectories: Using optimal alignment to reveal the structure of residential mobility. *Sociological methods & research* 32, 4 (2004), 559–598. http://journals.sagepub.com/doi/abs/10.1177/0049124103262683

Katherine Stovel, Michael Savage, and Peter Bearman. 1996. Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970. *Amer. J. Sociology* 102, 2 (1996), 358–399.

Matthias Studer. 2013. WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. (2013). https://archive-ouverte.unige.ch/unige:78576

Matthias Studer and Gilbert Ritschard. 2016. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179, 2 (2016), 481–511. http://onlinelibrary.wiley.com/doi/10.1111/rssa.12125/full

Matthias Studer, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2011. Discrepancy analysis of state sequences. *Sociological Methods & Research* 40, 3 (2011), 471–510. http://journals.sagepub.com/doi/abs/10.1177/0049124111415372

Matthias Studer, Emanuela Struffolino, and Anette E Fasang. 2018. Estimating the Relationship between Time-varying Covariates and Trajectories: The Sequence Analysis Multistate Model Procedure. *Sociological Methodology* (2018), 0081175017747122.

Warren S. Torgerson. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17, 4 (1952), 401–419.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)* 21, 1 (1974), 168–173. http://dl.acm.org/citation.cfm?id=321811

Eric D. Widmer and Gilbert Ritschard. 2009. The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research* 14, 1 (2009), 28–39. http://www.sciencedirect.com/science/article/pii/S1040260809000069

Paul Wiles. 2004. Policy and sociology. *The British journal of sociology* 55, 1 (2004), 31–34. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-4446.2004.00004.x/full