Data Quality Challenges with Missing Values and Mixed Types in Joint Sequence Analysis

Alina Lazar
Department of Computer Science
and Information Systems
Youngstown State University
Youngstown, OH
e-mail: alazar@ysu.edu

Ling Jin, C. Anna Spurlock,
Annika Todd
Energy Analysis and Environmental
Impacts Division, LBNL, Berkeley, CA
e-mail: ljin@lbl.gov

Kesheng Wu, Alex Sim
Computational Research Division,
LBNL, Berkeley, CA
e-mail: kwu@lbl.gov

Abstract—**The goal of this paper is to investigate the impact of missing values in categorical time series sequences on common data analysis tasks. Being able to more effectively identify patterns in socio-demographic longitudinal data is an important component in a number of social science settings. However, performing fundamental analytical operations, such as clustering for grouping these data based on similarity patterns, is challenging due to the categorical and multi-dimensional nature of the data, and their corruption by missing and inconsistent values. To study these data quality issues, we employ longitudinal sequence data representations, a similarity measure designed for categorical and longitudinal data, together with state-of-the art clustering methodologies reliant on hierarchical algorithms. The key to quantifying the similarity and difference among data records is a distance metric. Given the categorical nature of our data, we employ an "edit" type 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 similarity measure. Between variables with binary values and those with multiple nominal values, we find that the ability to overcome missing data problems is harder in the nominal domain versus the binary domain. Additionally, artificial clusters introduced by the alignment of leading missing values can be resolved by tuning the missing value substitution cost parameter.**

Keywords-joint sequence analysis; optimal matching; missing values; time series clustering; data quality

I. INTRODUCTION

Time series clustering plays an important role in temporal data mining research [1]. 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 word survey data with prominent and not commonly addressed data quality issues. We assess the effects of data quality issues by comparing clustering solutions from a systematically designed set of distance measures. 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 "similarity" measures and therefore change the clusters derived. We anticipate that similar challenges are present in working with 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 app, or other similar long-term tracking and categorization exercises using real world data.

In social science, time series clustering is used to study the spans of life trajectories in the form of sequence analysis [2]–[4]. By analyzing long-term life trajectory dynamics based on demographic characteristics, education and other lifestyle variables, it is possible to discover representative patterns from the overall life trajectory of a given individual's characteristics and the pathway through which one arrives at a given state, decision or behavior.

Traditional sequence analysis and new method development or evaluation have been heretofore focused on single variable trajectories [5]–[9]. This approach allows for relatively easy evaluation of missing observations that affects many real-world data sequences. However given the many factors and their interdependences that affect life trajectories, such as family planning, education, or employment, joint sequence analysis (formalized by [10], [11]) represents a more appropriate method. Its power relies on its capability to differentiate longitudinal experiences represented by multiple variables and therefore account more realistically for the inherent complexity of these types of problems [12]–[14].

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 longitudinal 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. 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 "perfect" survey data, arises from sequence alignment by development time (e.g., age) instead of calendar time. 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 both the beginning and ending of the sequences. For simplicity, this type of missing data will be referred to as **"alignment missing."** Both types of missing values are often encountered in real-world data, however their effects are not adequately evaluated especially in the joint sequence analysis literature.

Short internal survey gaps 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 [15], [16]. Using a single life course variable (employment), [16] illustrated the benefit of multiple imputation of missing values in minimizing biases in clustering.

For "alignment missing" and long survey gaps that consist of more contiguous missing values, the amount of information in the sequence is often not enough to impute the gaps. These missing values are usually included as a special category in computing distances between individual sequences [16]. According to review by Aisenbrey and Fasang [17], 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," will change or bias clustering results as well as bias mitigation methods.

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 interpretation as illustrated in [18]. 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 interpretation. Such interactive effects from both missing values and mixed data types have yet to be examined in the literature.

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 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; and (3) highlighting the interacting patterns between missing value treatments and mixed data types.

The rest of the paper is organized as follows. Section II describes the real-world data we use for the tasks. Section III explains the methods we use, including the distance measures, the systematic experimental clustering design, and the comparison metrics we use. Section IV evaluates clustering results and their dependence on missing value treatments and mixed data types. Section V concludes.

II. Data Description and Preprocessing

We analyzed data extracted from the Panel Study of Income Dynamics (PSID) [19] which includes a rich set of demographic and socio-economic information repeatedly collected from a large population available for years 1968 through 2015. The initial dataset contained over 17,000 records for individual people. 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. 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. Besides age, for the 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 education and marriage information). 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. This was done because we were interested in identifying 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 (year 1968) miss the leading part of the sequences ("alignment missing"). 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.

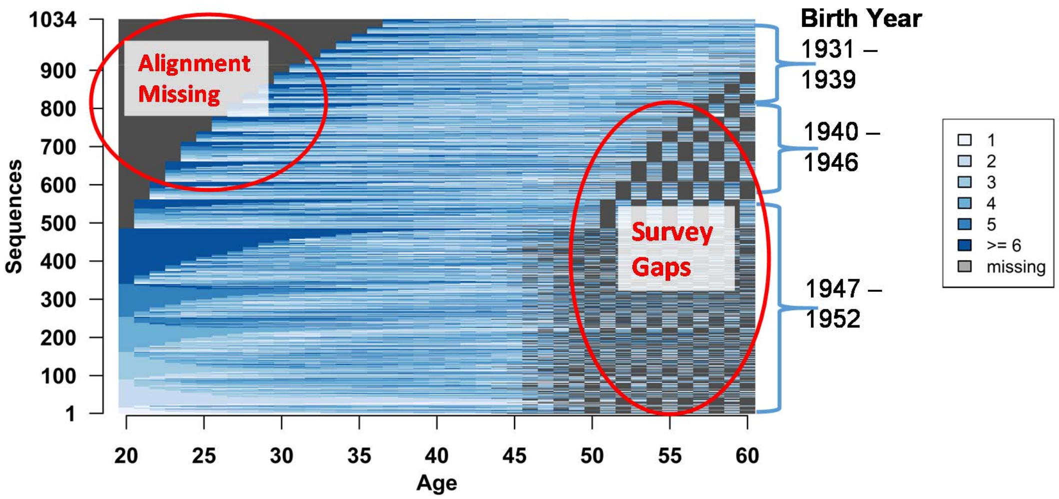

Figure 1. A plot family size sequence of all the individuals, to illustrate the missing value patterns that arise from survey gaps and missing segments after alignment by age.

III. Method

A. Distance Measures for Clustering

The notion of clustering hinges on the notion of distance, and therefore the concept and quantification of similarity is important for time-series data clustering. Metric distances such as the Euclidean distance fit well as similarity measures and are applied widely to identify patterns in longitudinal numeric data represented by time series [20]–[23]. 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 [24], [25], the *edit based distance 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 [26] and also in natural language processing for approximate character string matching [27]. 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 distances described above 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 similarity matrix has to take into consideration the contribution of each included variable. For this procedure, the indel and substitution costs are determined and set independently for each individual variable. Next, the substitution costs for the multivariate distance is calculated by averaging the substitution costs for the individual variables. The joint analysis approach follows [10] and is performed using the R package TraMineR version 2.0-6 [28].

B. Distance Experiments

In order 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.

(1) 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).

(2) "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. Note our approach is different from existing literature, where indel cost was adjusted for un-equal length sequences [29]. Tuning missing value substitution cost here provides the flexibility of adjusting distance contribution from missing values without affecting the transformation (indel or substitution) taken for non-missing values.

The "alignment missing," if applicable, is always present and therefore a substitution cost for the missing values (referred to as "**NA cost**" hereafter) must inevitably be chosen. The default "NA cost" is set to 2 in OM. 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:

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

To construct complete experiments, we apply these 4 missing value treatment cases to three types of joint sequence data: (1) binary (employment, marriage, education); (2) nominal (children under 8, and total family size); and (3) both of the above combined. The full 12 combinatorials allow for systematic comparison of the effects of both missing values and mixed data types in relation to each other on life course trajectory clustering.

C. Comparison Metrics

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 [30] 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 [31] suggest using Pearson’s correlation to evaluate and compare cluster solutions, which is an internal measure also known as “Point Biserial Correlation” (PBC equation below). 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_{dbin} be the standard deviation of d and d_{bin} respectively, and $s_{\mathrm{d,dbin}}$ 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 (ASW equation below) validates clustering performance based on the pairwise difference of between- and within-cluster distances. Originally proposed by Kaufman and Rousseeuw [32], 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 C_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.

2) Mutual Information Between two Clustering Solutions

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 done in [16]. 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 (5) and (6).

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

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

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

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

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

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

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

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

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

IV. Results and Discussions

A. Number of Clusters

Following previous research by [10] and [18] all the experiments were run use 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 III.B. 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 ($\boldsymbol{D_{binary}}$), followed by the combined domain ($\boldsymbol{D_{combined}}$) and ending with the nominal data domain ($\boldsymbol{D_{nominal}}$).

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 zero. 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.

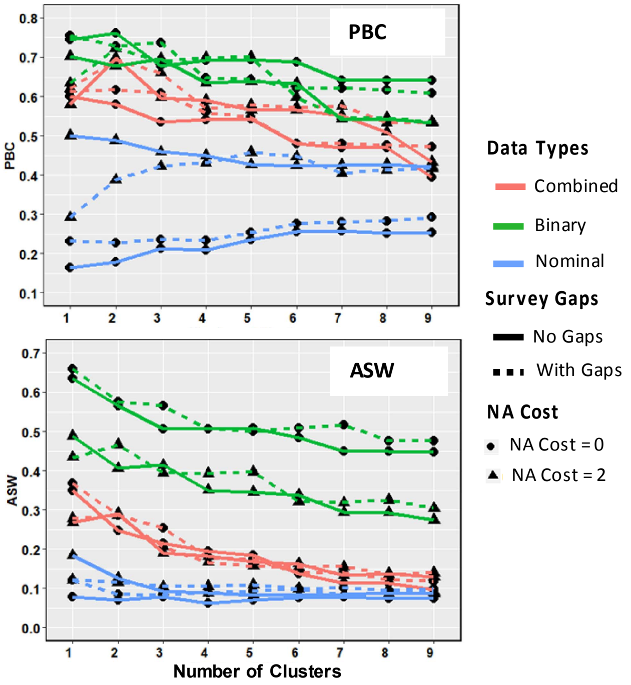

Figure 2. PBC and AWS as a function of number of clusters.

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 four clusters.

B. 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).

"Gap vs No Gap" (first two rows in Table I) 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 $|_{NA\_cost=0}$" 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 $\boldsymbol{D_{nominal}}$ much more than in $\boldsymbol{D_{binary}}$ or in $\boldsymbol{D_{combined}}$. Imputation effects in the $\boldsymbol{D_{nominal}}$ are also sensitive to the influence of alignment missing. In $\boldsymbol{D_{nominal}}$, clustering solutions using data with and without survey gaps become very little in common when NA cost changes from 2 to 0.

"NA cost = 2 vs NA cost = 0" (bottom two rows in Table I) represents the commonality between the choices of NA cost specification when missing values are included as a special category. In Table I, 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 $\boldsymbol{D_{nominal}}$ the most, especially when survey gaps are also present. $\boldsymbol{D_{combined}}$ also becomes more sensitive to the choice of NA cost when survey gaps are present.

TABLE I. NORMALIZED MUTUAL INFORMATION (*NMI*) BETWEEN CLUSTERING SOLUTIONS DERIVED WITH DIFFERENT MISSING DATA TREATMENTS (<0.5 VALUES ARE MASKED IN GREY)

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

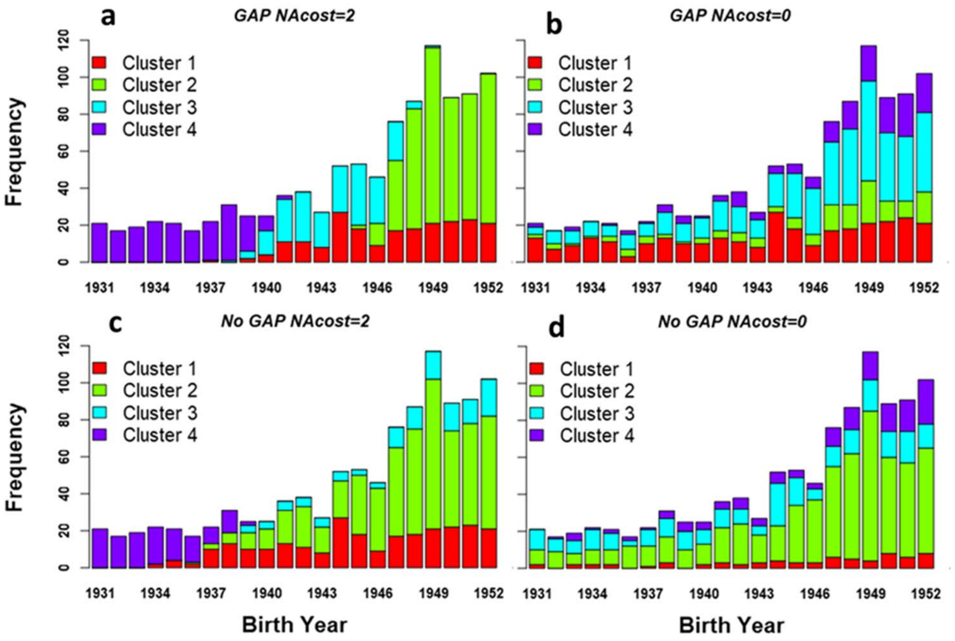

Figure 3. Birth year distribution of cluster assignments derived from $D_{combined}$ under four missing value treatments.

When NA cost is 2, the contribution of survey gaps and alignment missing to distance measures are maximized, which systematically biases upward the distance between sequences with and without missing segments. To explore the pattern of biases further, we examine the birth year distribution of clustering solutions under 4 cases of missing value treatments. 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. Figure 3 shows the birth year distribution of the 4-cluster solutions derived from $D_{combined}$. 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 (Figure 3a) are driven by age cohorts born 1947-1952, 1940-1946, and 1931-1939, respectively. These three age cohorts are subject to varying degrees of both alignment missing and survey gaps (Figure 1). Effects from both types of missing data are maximized at 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 (Figure 3c) 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 (Figure 3b and 3d) as "NA cost = 0" minimizes the contribution from missing values to the distance computation.

Clustering solutions in the $D_{nominal}$ are also subject to the age cohort induced biases and present similar artificial clusters (similar to Figure 3 and hence are not shown).

C. Effects of Mixed Data Types on Clustering

We have seen in section IV.A that distance matrices derived from the binary data domain (D_{binary}) are easier to cluster than those from mixed ($D_{combined}$) or nominal variables ($D_{nominal}$). In joint sequence analysis with mixed data types, the clustering solution from a combined data domain may favor the contributing domain that is easier to cluster (e.g., binary variables in our case) [18]. The interpretation of the clusters will then become problematic as they are not equally representative of all relevant domains.

Joint sequence analysis is found mostly useful when the individual domains are associated or interdependent [11]. However, missing values can complicate these domain association patterns. As we have seen from section IV.B, some treatments of missing values can create an artificial pattern in distance measures shared by all domains, leading to an "apparent" association. 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, and (2) how the commonality pattern changes with missing value treatments.

Figure 4 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. This is consistent with [18]. 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 (Table II).

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 4). 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 lighter blue first row of plots relative to the second row in Figure 4). Accordingly, the 4-cluster solution derived from $D_{combined}$ "appears" to perform better (measured by PBC and AWS) in the "NA cost =2" case than in the "NA cost = 0" case in the nominal domain $D_{nominal}$ (Table II). However, as we have seen in section IV.B, "NA cost = 2" maximizes the missing value contribution leading to a prominent pattern shared by dissimilarity matrices of all domains. Such shared biases in distance measures lead to an artificial association between the domains and thus increases the commonality in clustering solutions. 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.

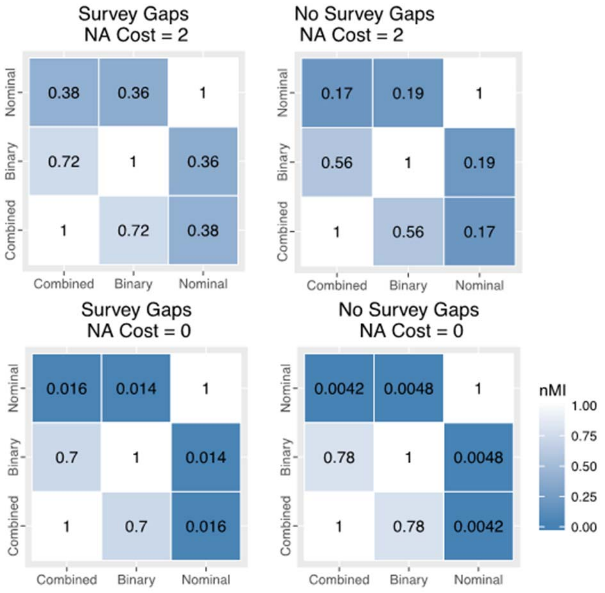

Figure 4. Normalized Mutual Information (nMI) between clustering solution derived from binary, nominal, and combined domains. Darker blue indicates greater differences between the pair.

TABLE II. PERFORMANCE OF THE 4-CLUSTER SOLUTION OF $D_{COMBINED}$ ON ITS CONTRIBUTING DOMAINS (D_{BINARY} AND $D_{NOMINAL}$) UNDER DIFFERENT MISSING VALUE TREATMENTS

Contr. 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

V. CONCLUSIONS

This paper evaluates two issues facing joint sequence analysis: missing values and mixed data types. Changes in clustering solutions are systematically assessed in a full combinatorial of experiment 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.

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. 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). 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. This study illustrates an extreme case that minimizes NA cost (=0), which eliminates the artificial cluster from alignment missing. Further study is needed to identify a case-by-case NA cost value that is most appropriate. An alternative practice could be to conduct clustering separately for different age cohorts to avoid serious biases in distance measures and clustering solution introduced by alignment missing, though this may not always be appropriate depending on the goals of the analysis being undertaken.

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 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. In comparison, when missing value effects are eliminated by setting NA cost to 0, there is a reduction in the disproportionate association between the combined domain and nominal domain. This finding highlights the importance of the choice of missing value treatment in correct interpretation of the clustering solutions represented in the contributing domains.

The analysis framework illustrated here can be easily extended to other variants of OM and cost definitions so that the method dependence of our conclusions can be evaluated further.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and 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

[1] D. Kotsakos, G. Trajcevski, D. Gunopulos, and C. C. Aggarwal, *Time-Series Data Clustering*. 2013.

[2] H. Bras, A. C. Liefbroer, and C. H. Elzinga, “Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900,” *Demography*, vol. 47, no. 4, pp. 1013–1034, 2010.

[3] E. D. Widmer and G. Ritschard, “The de-standardization of the life course: Are men and women equal?,” *Adv. Life Course Res.*, vol. 14, no. 1, pp. 28–39, 2009.

[4] R. Schumacher, K. Matthijs, and S. Moreels, “Migration and reproduction in an urbanizing context. a sequence analysis of family life courses in 19th century Antwerp and Geneva,” 2012.

[5] M. Studer, G. Ritschard, A. Gabadinho, and N. S. Müller, “Discrepancy analysis of state sequences,” *Sociol. Methods Res.*, vol. 40, no. 3, pp. 471–510, 2011.

[6] C. H. Elzinga, “Sequence analysis: Metric representations of categorical time series,” *Sociol. Methods Res.*, 2006.

[7] J.-A. Gauthier, E. D. Widmer, P. Bucher, and C. Notredame, “How much does it cost? Optimization of costs in sequence analysis of social science data,” *Sociol. Methods Res.*, vol. 38, no. 1, pp. 197–231, 2009.

[8] B. Halpin, “Three Narratives of Sequence Analysis,” in *Advances in Sequence Analysis: Theory, Method, Applications*, Springer, Cham, 2014, pp. 75–103.

[9] M. Studer and G. Ritschard, “What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 179, no. 2, pp. 481–511, 2016.

[10] G. Pollock, “Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 170, no. 1, pp. 167–183, 2007.

[11] J.-A. Gauthier, E. D. Widmer, P. Bucher, and C. Notredame, “Multichannel sequence analysis applied to social science data,” *Sociol. Methodol.*, vol. 40, no. 1, pp. 1–38, 2010.

[12] P. Johnson, “Making social science useful,” *Br. J. Sociol.*, vol. 55, no. 1, pp. 23–30, 2004.

[13] H. Lauder, P. Brown, and A. H. Halsey, “Sociology and political arithmetic: some principles of a new policy science,” *Br. J. Sociol.*, vol. 55, no. 1, pp. 3–22, 2004.

[14] P. Wiles, “Policy and sociology,” *Br. J. Sociol.*, vol. 55, no. 1, pp. 31–34, 2004.

[15] P. Royston and others, “Multiple imputation of missing values,” *Stata J.*, vol. 4, no. 3, pp. 227–41, 2004.

[16] B. Halpin, “Multiple Imputation for Life-Course Sequence Data,” May 2012.

[17] S. Aisenbrey and A. E. Fasang, “New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course, *Sociological. Methods & Research.*, vol. 38, no. 3, pp. 420–462, 2010.

[18] R. Piccarreta, “Joint Sequence Analysis: Association and Clustering,” *Sociol. Methods Res.*, vol. 46, no. 2, pp. 252–287, 2017.

[19] *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)*.

[20] T. W. Liao, “Clustering of time series data—a survey,” *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[21] S. Rani and G. Sikka, “Recent techniques of clustering of time series data: a survey,” *Int. J. Comput. Appl.*, vol. 52, no. 15, 2012.

[22] T. Fu, “A review on time series data mining,” *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 164–181, 2011.

[23] L. Jin, D. Lee, A. Sim, S. Borgeson, K. Wu, C. A. Spurlock, and A. Todd. “Comparison of Clustering Techniques for Residential Energy Behavior using Smart Meter Data,” in *AI for Smart Grids and Buildings Workshop* at the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, 2017.

[24] A. Abbott and J. Forrest, “Optimal matching methods for historical sequences,” *J. Interdiscip. Hist.*, vol. 16, no. 3, pp. 471–494, 1986.

[25] A. Abbott and A. Hrycak, “Measuring resemblance in sequence data: An optimal matching analysis of musicians’ careers,” *Am. J. Sociol.*, vol. 96, no. 1, pp. 144–185, 1990.

[26] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.

[27] R. A. Wagner and M. J. Fischer, “The string-to-string correction problem,” *J. ACM JACM*, vol. 21, no. 1, pp. 168–173, 1974.

[28] A. Gabadinho, G. Ritschard, N. S. Mueller, and M. Studer, “Analyzing and visualizing state sequences in R with TraMineR,” *J. Stat. Softw.*, vol. 40, no. 4, pp. 1–37, 2011.

[29] K. Stovel and M. Bolan, “Residential trajectories: Using optimal alignment to reveal the structure of residential mobility,” *Sociol. Methods Res.*, vol. 32, no. 4, pp. 559–598, 2004.

[30] M. Studer, “WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R,” 2013.

[31] C. Hennig and T. F. Liao, “Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification,” Technical report, 2010.

[32] L. Kaufman and P. J. Rousseeuw, “Partitioning around medoids (program pam),” *Find. Groups Data Introd. Clust. Anal.*, pp. 68–125, 1990.