Handling Missing Values in Joint Sequence Analysis

Alexandra Ballow[1], Alina Lazar[1], Alex Sim[2], Kesheng Wu[2]

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

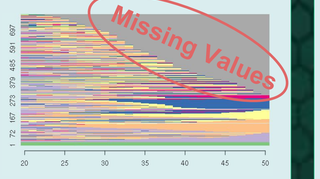
BERKELEY LAB | WORKFORCE DEVELOPMENT & EDUCATION

ABSTRACT

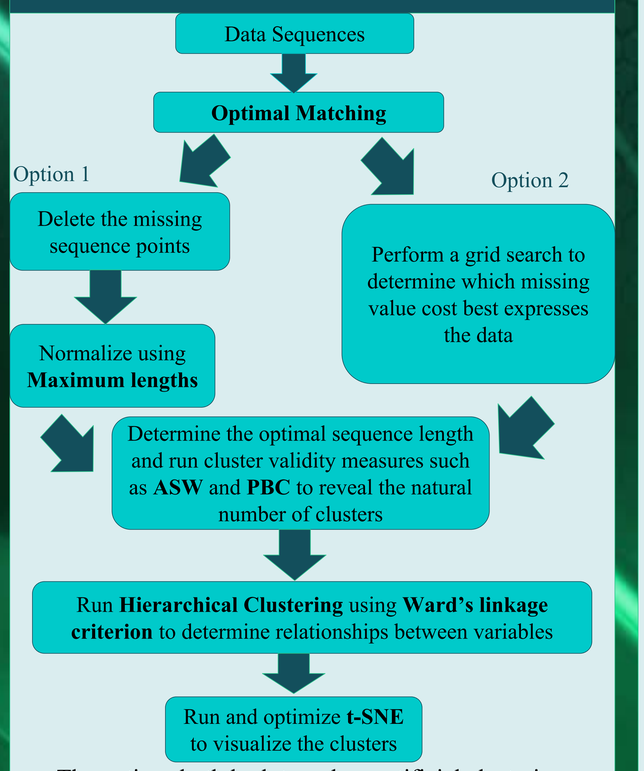
This study focuses on developing methodologies to minimize the effects of incomplete data. Specifically, it hopes to reduced the noise and bias caused in categorical sequence data by data gaps. Some strategies investigated include choosing a substitution "cost" to replace missing values and deleting the missing values at the end of a sequence. Cluster validity metrics are used to determine the accuracy of the unsupervised clustering algorithms and t-SNE is employed to visualize clusters and age biases. It became clear that deleting missing values provided the best results, but all data sets are different. Thus this study recommends employing the studied procedures before conducting analysis on longitudinal sequence data to ensure the results are unbiased. After these tests optimize the data, clustering is conducted to understand the correlation between a person's state in life and their travel.

DATA

- Data acquired from the WholeTraveler survey
- Consists of demographic and travel information from nearly a thousand Bay Area citizens
- This study focuses on the binary variables including:
  - Whether they walked or used a bicycle
  - Having a child
  - Having a partner
  - Use of public transport
  - Employment Status
- Is full of missing values, as demonstrated by the family size sequences (right)



METHODS



- The options both look to reduce artificial clustering by alignment variable (age).

RESEARCH QUESTION

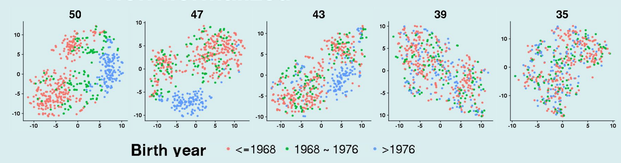
What effect does the treatment of missing values in categorical sequence data have on cluster analysis results?

RESULTS

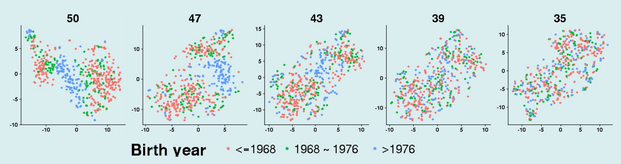
Normalization

~Option 1~

Non-normalized



Normalized

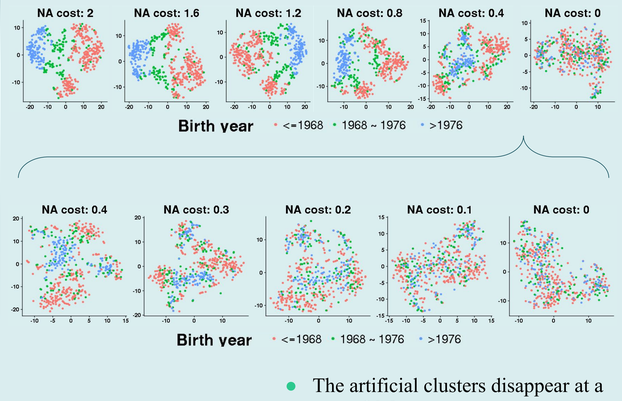


- The normalized data is chosen as the artificial clusters are less defined

Substitution Cost

~Option 2~

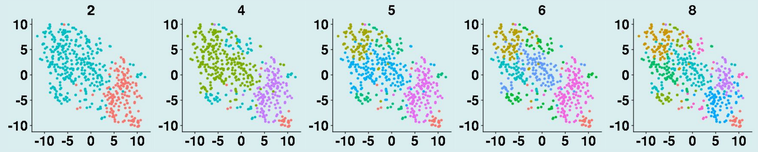
Varying Substitution Cost



- The artificial clusters disappear at a cost of 0.2

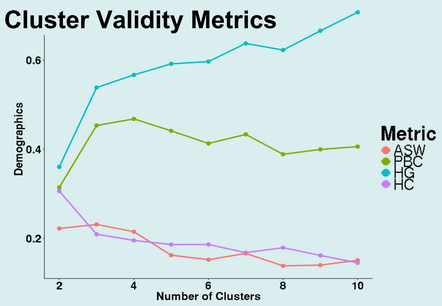
Determining Number of Clusters

Varying Number of Clusters



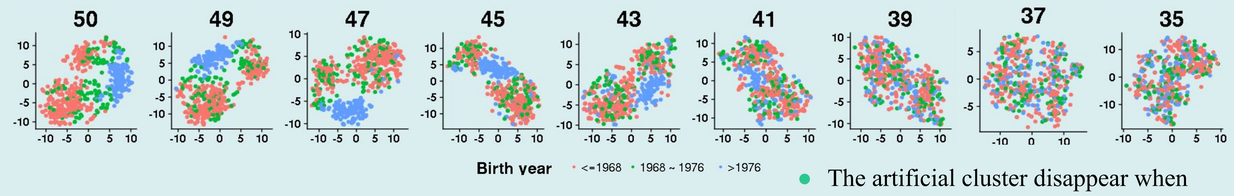
- These tests indicate that there are 4 natural clusters in this data as there are peaks in the cluster validity measures at 4 clusters

Cluster Validity Metrics



Determining the Length of Sequences

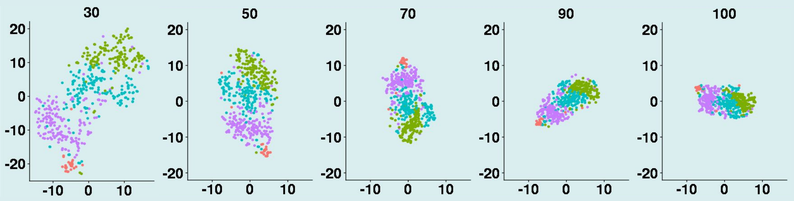
Varying Ending Age



- The artificial cluster disappear when the years 20 to 35 are included

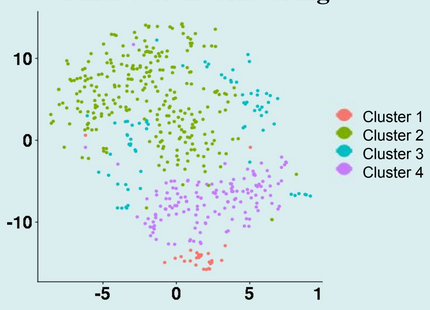
Clustering

Varying Perplexities



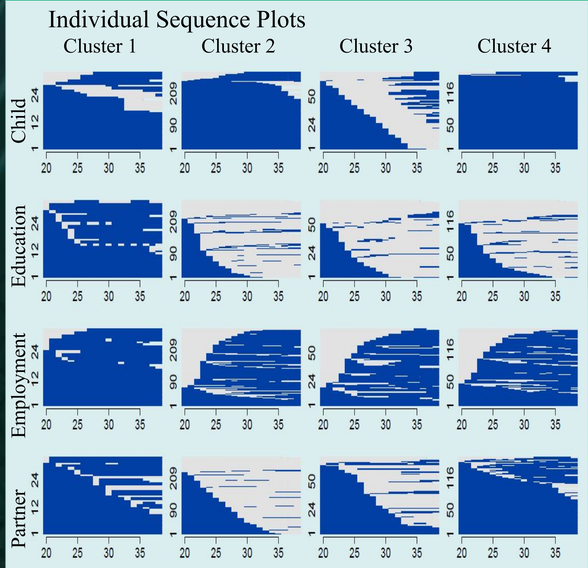
- After a careful evaluation of different perplexities, 50 is chosen due to its lack of noise and evaluation of outliers

Final t-SNE Clustering



FINAL CLUSTERS

Individual Sequence Plots



- Each of the clusters above represents a type of person in our data set
- Cluster 1 and 3 are made up of less people so those respondents followed a pattern atypical of the general survey pool

Not in that state
In that state

CONCLUSIONS

Recently released data detailing travel tendencies was wrought with large groups of missing values, typically found in longitudinal data. This project was intended to investigate and mitigate the effect of missing data on clustering results on this new data. After testing two different procedures to minimize the biases created by missing values, it is determined that removing missing values from the end of the sequences and normalizing the distance measure produced the best results. This procedure generated more interpretable results, with less parameter tuning and less error. Proceeding with the results acquired during this step, it was possible to create clusters and sequences plots which will be evaluated in future studies.

ACKNOWLEDGMENTS

This work was prepared in partial fulfillment of the requirements of the Berkeley Lab Undergraduate Faculty Fellowship (BLUFF) Program, managed by Workforce Development & Education at Berkeley Lab.
I want to thank Ling Jin, Annika Todd, and Anna Spurlock from the Energy Analysis Environmental Impacts Group for providing datasets and guidance for the project.

SDM | Youngstown State University | U.S. Department of Energy, Office of Science