Visualizing Sequences in the Social Sciences: Relative Frequency Sequence Plots

Tim F. Liao, University of Illinois

5 April 2013
Workshop in Methods, Indiana University
Outline

• Why visualization?
• Why sequence analysis?
• Data representation versus data summarization
• What’s wrong with the existing methods of sequence visualization?
• Relative frequency sequence plots
• Plotting variability
• Assessing representation
• Future directions

Acknowledgements: the presentation is based on a collaborative project with Anette Fasang, Humboldt University Berlin & Wissenschaftszentrum Berlin.

(1) Above all else show the data, i.e. tell the truth about the data

(2) maximize the data-ink ratio

(3) erase non-data-ink

(4) erase redundant data-ink

(5) revise and edit

(1) Give an overview

(2) convey scale and complexity

(3) explore the data and compare

(4) communicate information

(5) tell a story

(6) attract attention/stimulate interest
Visualization: Songwriters’ Ratings by Age

Simon

Wednesday Morning, 3 AM
One-Trick Pony
The Rhythm of the Saints
You’re the One
Surprise
So Beautiful or So What

McCartney

Abbey Road
Help!
Band on the Run
Wings Over America
Tug of War
Off the Ground
Flaming Pie
Driving Rain
Chaos and Creation In the Backyard
Memory Almost Full

Dylan

Bob Dylan
Nashville Skyline
Desire
Blood on the Tracks
Infidels
Oh Mercy
World Gone Wrong
Tempest
Together Through Life

Count of Albums
Sociological Rationale for Sequence Analysis

- Elder (1985): "Life patterns are structured by variations in the timing, duration, and order of events."
- Rindfuss, Swicegood, and Rosenfeld (1987) in “Disorder in the Life Course: How Often and Does it Matter?” analyzed NLS of class 72 data, 5 school-employment events of 8 years, if equal probability $5^8=390,625$ sequences. They encouraged researchers to take a more careful look at the life course as it is actually lived
- As such, the premises and the objects of sequences analysis are rather different from other types of longitudinal analysis
## Why Sequence Analysis?
A Comparison of Panel Regression, Event History and Sequence Analysis

<table>
<thead>
<tr>
<th>Theoretical concept</th>
<th>Event history analysis</th>
<th>Panel Regression</th>
<th>Sequence analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>transition</td>
<td>change</td>
<td>Trajectory, story, narrative</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scientific tradition / Theory</th>
<th>stochastic data modeling culture</th>
<th>narrative positivism</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Assumption about data generation</th>
<th>stochastic process, causality</th>
<th>none</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Objective to identify…</th>
<th>probability of single transitions/durations</th>
<th>probability change</th>
<th>patterns of sequential equivalence</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Broader theoretical implications</th>
<th>Structure ==&gt; occurrence of / change in events</th>
<th>Structure + agency interact across process ==&gt; sequence</th>
</tr>
</thead>
</table>

→ Complementary methods for different questions
<table>
<thead>
<tr>
<th>Type</th>
<th>Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable-Based</td>
<td>Probability of Dependent Variable</td>
</tr>
<tr>
<td></td>
<td>Timing of DV increasing/happening</td>
</tr>
<tr>
<td></td>
<td>Independent variables causally explain DV</td>
</tr>
<tr>
<td></td>
<td>(Event History, Panel Models, Growth curve modeling…)</td>
</tr>
<tr>
<td>Process-Based</td>
<td>Thick Description</td>
</tr>
<tr>
<td></td>
<td>Open, patterns of states (no DV)</td>
</tr>
<tr>
<td></td>
<td>Multivariable Regularities</td>
</tr>
<tr>
<td></td>
<td>(Sequence Analysis, Optimal Matching, Hamming …)</td>
</tr>
</tbody>
</table>
Current sequence graphs broadly fall into two groups:

1. **Representation**: the sequence index plot (and its extensions): primarily information about *qualitative* sequence patterns, also show variation *within* individual sequences over time

2. **Summarization**: plots that aggregate and summarize sequence information: quantitative information about sequence characteristics, state distributions, variation *between* sequences (many developed within TraMineR package, Gabadinho et al. 2011)
Example Data: German Life History Study (GLHS)

- The larger study collected data since early 1980s on the life histories of some 8,500 men and women from 20 selected birth cohorts in West Germany and of more than 2,900 men and women from 13 selected birth cohorts in East Germany.
- For this analysis: family formation sequences of 474 East (N=132) and West (N=342) German women born 1971.
- Retrospective life history data, collected in 1997/1998 & follow up in 2005; we only use the data available in 2005, thus we can follow family formation until age 34.
- Age range: 15-34
- Seven states of family formation: single, in a relationship, cohabiting, cohabiting with a child, married, married with a child, and divorced/widowed.
Data Representation Graphs: the Sequence Index Plot and Extensions
Data Summarization Graphs: Quantitative Sequence Information

(a) State Distribution Plot

(b) KM Survival Curves

(c) Modal State Plot

(d) Mean Time Plot

- S
- R
- CNC
- CC
- MNC
- MC
- DW

Family formation states

Modal state sequence (0 occurrences, freq=0%)
Visualization of Sequences: Sequence Index Plots

- Introduced by Scherer 2001
- Stata.ado by Kohler, Brzinsky-Fay and Luniak 2006
- Each sequence represented by one line across a process time axis, different states indicated by different colors
Alternative Visual Methods

• Tree representation of clusters of transitions to adulthood (Aassve et al. 2007)

• MDS Sequence index plot of family formation trajectories (Piccarreta & Lior 2010)
Mean Time Spent in Each state among East and West German Women

**East Germany**

<table>
<thead>
<tr>
<th>State</th>
<th>Cohabit, Child</th>
<th>Cohabit, No Child</th>
<th>Divorced/Widowed</th>
<th>In Relationship</th>
<th>Married, Child</th>
<th>Married, No Child</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CN</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DW</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MN</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**West Germany**

<table>
<thead>
<tr>
<th>State</th>
<th>Cohabit, Child</th>
<th>Cohabit, No Child</th>
<th>Divorced/Widowed</th>
<th>In Relationship</th>
<th>Married, Child</th>
<th>Married, No Child</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CN</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DW</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MN</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>43</td>
<td>86</td>
<td>130</td>
<td>173</td>
<td>216</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend:**
- **Cohabit, child**
- **Cohabit, no child**
- **Divorced/widowed**
- **In a relationship**
- **Married, child**
- **Married, no child**
- **Single**
Advantages of Sequence Index Plots

- Holistic representation of processes
- Taking individual sequence seriously as a unit of analysis
- Convey large amounts of information in single graph: prevalence and timing of states, intermediary states, overall variability within and between sequences etc...
Problems of the Basic Sequence Index Plot

Sequence index plots can be deceptive due to the following reasons:

1. **Over-plotting:** deceptive in terms of relative importance of individual sequences, problematic for N larger than a couple of hundreds

2. **Unfair Comparison:** difficult to compare groups of unequal sizes

3. **Reliance on Order:** lose sight of central tendencies in the timing of more than one transition within sequences
Problem I: Over-Plotting, 474 Cases
Problem I: Over-Plotting, 948 Cases
Problem I: Over-Plotting, 1932 Cases
Problem II: Comparative Difficulty
Relative Frequency Sequence Plots

• To deal with over-plotting of large $N$: we draw random or representative sequences (as two different approaches) and present them as relative frequency sequence plots

• For comparing groups of unequal sizes: we compare differences in relative frequency (%) of specific sequences (and states they contain) across plots of unequal sizes
Relative Sequence Plots of Family Formation Sequences in West and East Germany

→ width of each line represents relative frequency of sequence in subgroup

Relatively more cohabitation with or without child in East Germany than in West Germany
Sequence Index Plots of Clusters of Family Formation (based on Lesnard’s dynamic OM distance)
Relative Frequency Sequence Plots of Family Formation Clusters (sum to 100%)
Relative Sequence Plots for Cohort Comparison

Cohort 1953/55, N = 731
Cohort 1971, N = 474

Early marriage & child
Cohabitation, late marriage & child
Delay

percent

15 20 25 30
age

single
cohabiting, no child
cohabiting, child
married, no child
married, child
divorced/widowed
Problem III: Order of Sequences

- One may lose sight of central tendencies in multiple transitions within sequences
- Ordering of sequences can change one’s overall perception of a graph
Ordered by Timing of Cohabitation or Marriage?

Deceptive/not informative on timing/prevalence of marriage when ordered by cohabitation and vice versa.
Multiple Transition Curve Plots Based on Kaplan Mayer Survival Curves

\[ \hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}. \]

Calculate separate transition curves for each transition of interest:

1. in a relationship
2. cohabitation
3. marriage
4. cohabitation with child
5. marriage with child
The Relative Frequency Plots for East & West Germany
Multiple Transition Curves for Family Formation Clusters
The Relative Frequency Plot for the Marriage/Late Child Cluster
The Relative Frequency Plot for the Marriage/Early Child Cluster
The Relative Frequency Plot for the Late Marriage Cluster
Are They Different Enough?
Are They Different Enough?

There should be no significant difference between 2 random 100 sequences.
Procedure for Constructing Relative Frequency Sequence Plots

1. Sort the sequences according to a substantively meaningful principle.
2. Divide the sorted sample into $K$ groups having the same frequency.
3. Choose the medoid sequence of each frequency group to represent this group.
4. Plot the selected representative sequences as a sequence index plot.
5. Plot the dissimilarities to the medoid within each frequency group as a box and whisker plot.
6. Calculate an $R^2$ statistic and an $F$-test to evaluate the goodness of fit and compare different specifications of the relative frequency sequence plot.
Sorting Principles

1. The timing of a focal transition of interest, e.g., age of first marriage.

2. Scores on an MDS factor.

3. Dissimilarities to the most frequent sequence.

4. Dissimilarities to a theoretically ideal-typical sequence.
Calculating the $R^2$ statistic and $F$-test to Evaluate the Goodness of Fit

1. Let $S_i$ denote the relative distance of a relative frequency group:

   $S_i = \frac{\sum_{j=1}^{n_j} q_{ij}^2}{\sum_{j=1}^{n_j} Q_{ij}^2}$

2. $R^2$ statistics can be defined as:

   $R^2 = 1 - \sum_{i=1}^{K} S_i = 1 - \sum_{i=1}^{K} \frac{\sum_{j=1}^{n_j} q_{ij}^2}{\sum_{j=1}^{n_j} Q_{ij}^2}$

3. $F$-test for assessing the sufficiency of $K$ relative frequency groups to capture heterogeneity:

   $F = \frac{ESD}{USD} = \frac{R^2 / (K - 1)}{(1 - R^2) / (N - K)}$
Sequence index plot, worst and best specification of relative frequency (RF) sequence plot and dissimilarity-to-medoid plot of women’s family formation based on $R^2$ and $F$-test in table 2 with $K=52$
Sequence index plot, relative frequency sequence plot, and dissimilarity-to-medoid plot of family formation in East and West Germany.
Future Directions

1. Making visualization dynamic and interactive
2. Assessing and testing group differences
3. Building causal models
Thank you!
Descriptive information about the example data on family formation of women born 1971 in East and West Germany between age 15 and 33 (N=474)

<table>
<thead>
<tr>
<th>Sequence state</th>
<th>Percent women ever experienced the event</th>
<th>Mean age at first occurrence of the event</th>
<th>Mean duration in event (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single (S)</td>
<td>97.68</td>
<td></td>
<td>62.06</td>
</tr>
<tr>
<td>In a relationship (R)</td>
<td>98.31</td>
<td>18.71</td>
<td>53.20</td>
</tr>
<tr>
<td>Cohabiting, no child (CNC)</td>
<td>77.00</td>
<td>22.97</td>
<td>29.52</td>
</tr>
<tr>
<td>Cohabiting, child (CC)</td>
<td>18.35</td>
<td>25.90</td>
<td>8.69</td>
</tr>
<tr>
<td>Married, no child (MNC)</td>
<td>61.81</td>
<td>25.07</td>
<td>17.90</td>
</tr>
<tr>
<td>Married, child (MC)</td>
<td>58.86</td>
<td>26.48</td>
<td>43.89</td>
</tr>
<tr>
<td>Divorced/widowed (DW)</td>
<td>4.01</td>
<td>27.29</td>
<td>.75</td>
</tr>
</tbody>
</table>

Source: German Life History Study (GLHS)
Overview of $R^2$ and $F$-test for different combinations of sorting principle, distance measure, and $K$ frequency groups, for example data on family formation with $N=474$

<table>
<thead>
<tr>
<th>$K$</th>
<th>Sorting principle</th>
<th>Distance measure</th>
<th>$R^2$</th>
<th>$F$</th>
<th>Upper limit for $\alpha=0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>Age of first marriage</td>
<td>DHD</td>
<td>.450</td>
<td>1.650</td>
<td>1.320</td>
</tr>
<tr>
<td></td>
<td>Age of first marriage</td>
<td>OM</td>
<td>.406</td>
<td>1.374</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of first marriage</td>
<td>LCP</td>
<td>.282</td>
<td>.792</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>DHD</td>
<td>.495</td>
<td>1.972</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>OM</td>
<td>.491</td>
<td>1.945</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>LCP</td>
<td>.409</td>
<td>1.391</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>Age of first marriage</td>
<td>DHD</td>
<td>.333</td>
<td>2.525</td>
<td>1.441</td>
</tr>
<tr>
<td></td>
<td>Age of first marriage</td>
<td>OM</td>
<td>.294</td>
<td>2.110</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of first marriage</td>
<td>LCP</td>
<td>.122</td>
<td>.706</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>DHD</td>
<td>.369</td>
<td>2.961</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>OM</td>
<td>.401</td>
<td>3.386</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>LCP</td>
<td>.256</td>
<td>1.747</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Age of first marriage</td>
<td>DHD</td>
<td>.302</td>
<td>3.582</td>
<td>1.566</td>
</tr>
<tr>
<td></td>
<td>Age of first marriage</td>
<td>OM</td>
<td>.263</td>
<td>2.954</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age of first marriage</td>
<td>LCP</td>
<td>.072</td>
<td>.644</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>DHD</td>
<td>.339</td>
<td>4.250</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>OM</td>
<td>.367</td>
<td>4.789</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDS score</td>
<td>LCP</td>
<td>.195</td>
<td>2.005</td>
<td></td>
</tr>
</tbody>
</table>

Note: DHD = Dynamic Hamming Distance, OM = Optimal Matching with substitution costs = 2 and indel costs=1, LCP = Longest Common Prefix distance.