# SEQUENCE AND CLUSTER ANALYSIS: EXPLANATIONS AND APPLICATIONS

**Amal Harrati, PhD**

**Mathematica**

**Oct. 20, 2021**

# Agenda

1) Goals and application for sequence and cluster analysis

2) Explanation of mechanics

3) Responses to readings, QA, discussion

# When to use sequence or clustering analysis?

- Long trajectories of data that are different across unit of observation.
- Lifecourse analysis; time-use data; genetic variations
- Data reduction technique

# Goal of sequence analysis

1) **Comparison:** Using distance measures (ie. optimal matching, Hamming, etc.)

2) **Grouping:** Similar sequences are grouped using cluster analysis or multidimensional scaling

3) **Application**: Use the grouped sequences are dependent or independent variables in standard statistical analysis (ie. regression, survival analysis, etc.)

Source: Brzinsky-Fay and Kohler, 2006

# Step 1: Preparation of the Data

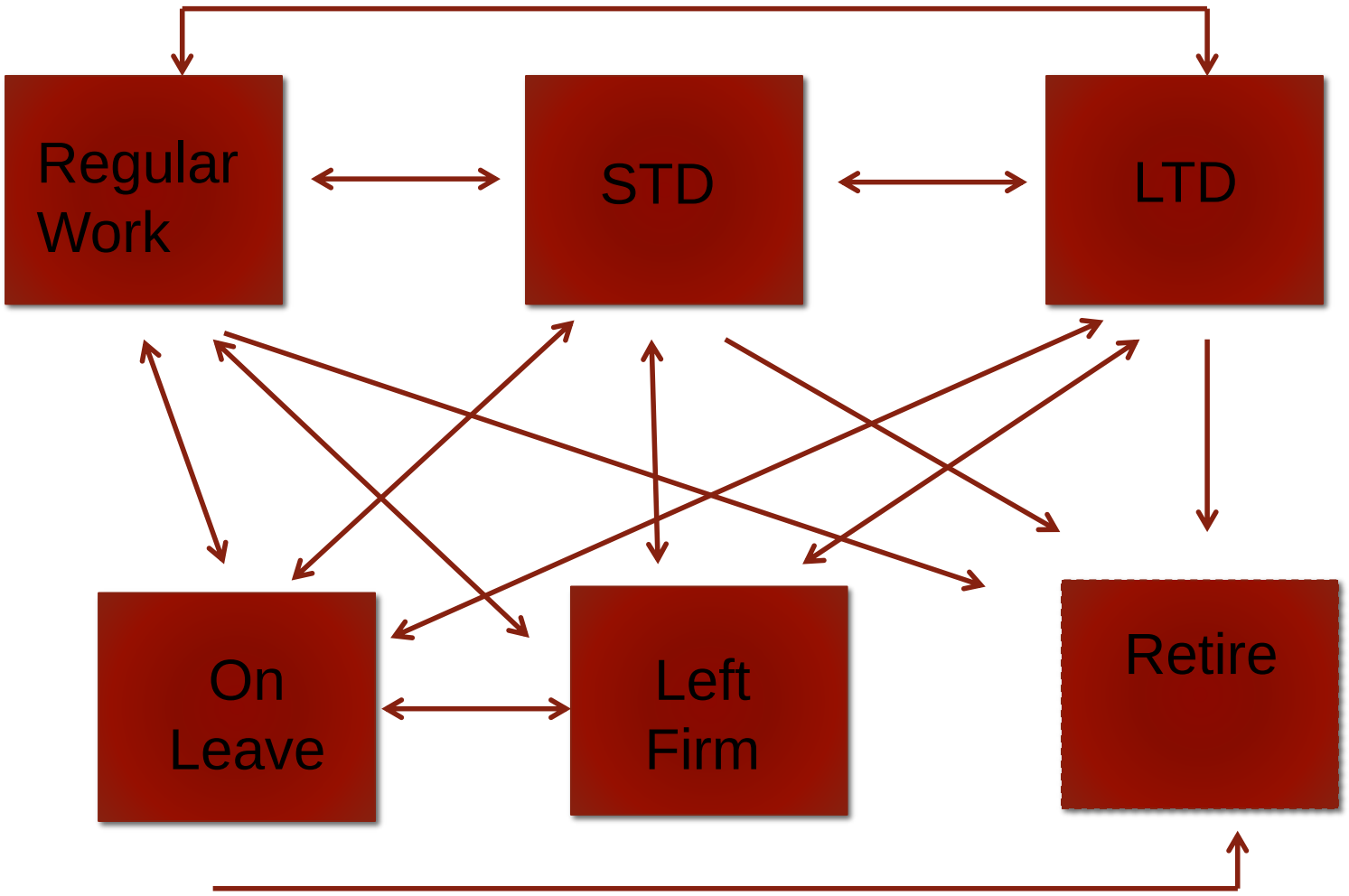
1) Consider your research question
   SA should consider not just timing, but *sequencing* and *duration*. It is largely intended to capture *trajectories, or sequences of linked states*

2) Is this a good application of the data?
   Is there enough variation in the sequence?
   Is there a long enough time period?

3) Creating your states

4) Consider temporality
   (Elzinga vs. Halpin)

# Example of a “state”



# Step 2: Create a difference metric

1) Optimal matching

2) Elzinga's Longest Common Subsequence

Source: Studart and Ritschard, 2016

# Optimal Matching

Three operations:

1) Insertion: one state is inserted into a sequence

2) Deletion: one state is deleted from a sequence

3) Substitution: one state is replaced by another state to into the sequence

Must assign a cost to each operation, and each suboperation.

Algorithm makes pairwise costs for all sequences and the distance is the minimal cost.

# How to compare two distinct sequences

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |

Slide source: Ivan Mejia Guevara

# How to compare two distinct sequences

| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |

- Alignment 1. Use substitution operations only:

| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |
| | 0 | 0 | s | s | 0 | s | 0 | s | 0 | 0 | 0 | 0 |

Total cost = 4s

Slide source: Ivan Mejia Guevara

# How to compare two distinct sequences

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |
| | 0 | 0 | s | s | 0 | s | 0 | s | 0 | 0 | 0 | 0 |

Slide source: Ivan Mejia Guevara

# How to compare two distinct sequences

- Alignment 2. Alternatively, using insertion/deletion and substitutions:

| Sequence 1 | A | A | A | | B | B | C | C | B | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B | |
| | 0 | 0 | s | d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | d |

Total Cost = s + 2D

Slide source: Ivan Mejia Guevara

# How to compare two distinct sequences

| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |

- Alignment 1. Use substitution operations only:

| Sequence 1 | A | A | A | B | B | C | C | B | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B |
| | 0 | 0 | s | s | 0 | s | 0 | s | 0 | 0 | 0 | 0 |

Total cost = 4s

- Alignment 2. Alternatively, using insertion/deletion and substitutions:

| Sequence 1 | A | A | A | | B | B | C | C | B | B | B | B | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sequence 2 | A | A | B | C | B | B | C | C | B | B | B | B | |
| | 0 | 0 | s | d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | d |

Total cost = s + 2d

# Criticisms of Optimal Matching

- Operations imply a change in the time scale.
- If there is not a clear ranking in states, costs can be sociologically meaningless.
- Censoring creates unequal sequence length: shorter sequences have higher costs

# “Solutions to” or Alternatives to OM

Using observed transition matrix

Elzinga’s longest common subsequence

Censoring creates unequal sequence length: shorter sequences have higher costs

# Elzinga's longest common sequence

Using observed transition matrix

Elzinga's longest common subsequence

Censoring creates unequal sequence length: shorter sequences have higher costs

# Step 3: Clustering

Goal:

Provide a way of grouping sequences according to their distance metrics (similarities and dissimilarities to other sequences)

# Step 3: Clustering

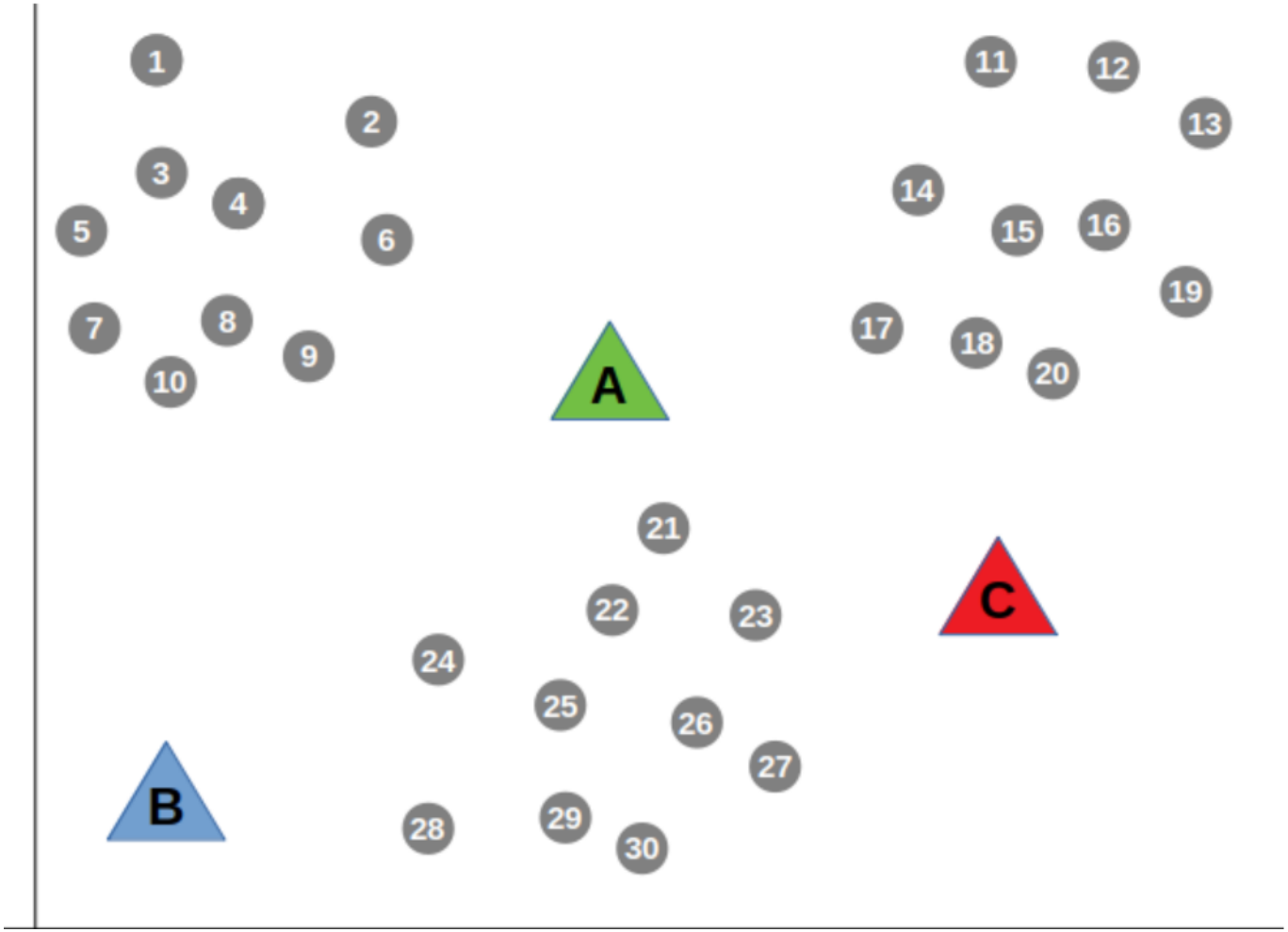
Different methods are available:

1) Ward's method
   *Minimizes total variance within clusters*

2) K-means
   *Creates centroids and minimizes mean values of centroids*

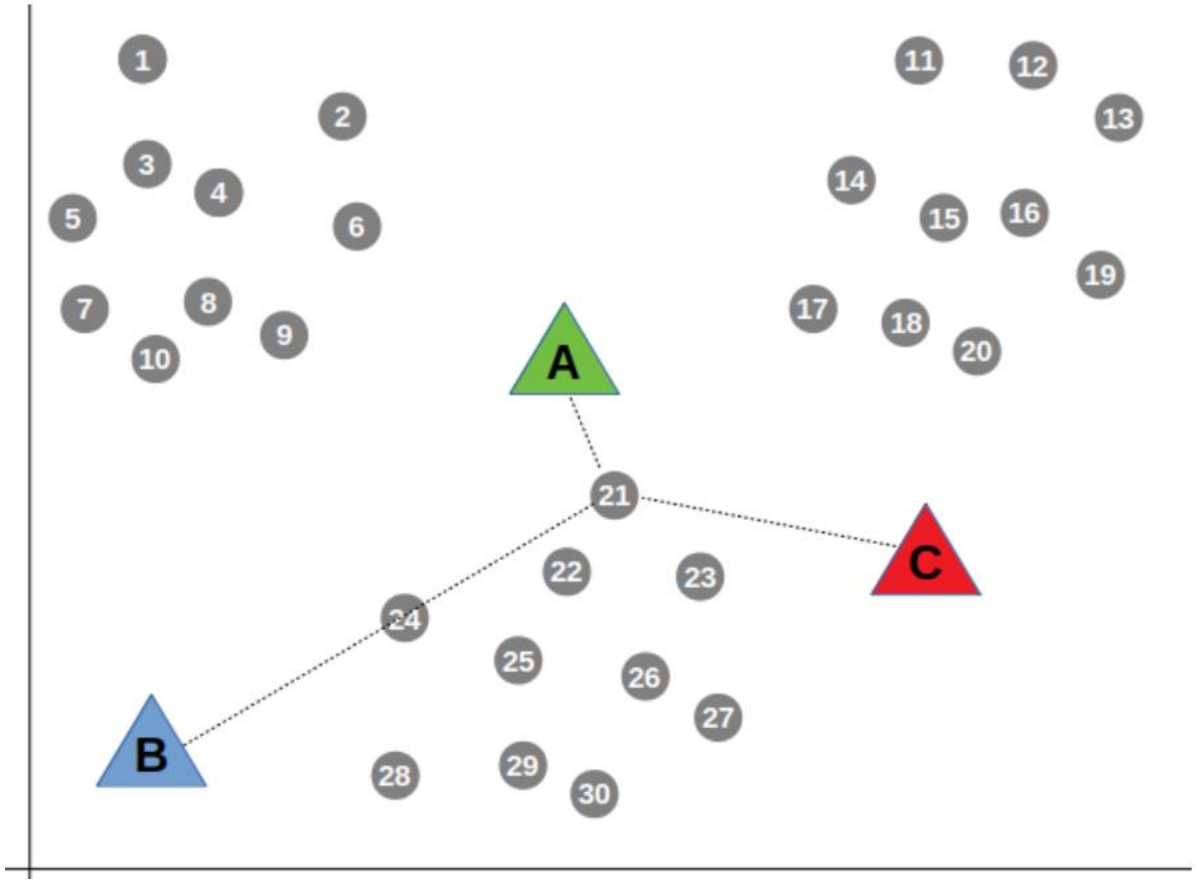
3) Partition around mediods
   Similar to k-means; splits data set into

# K-means or PAM



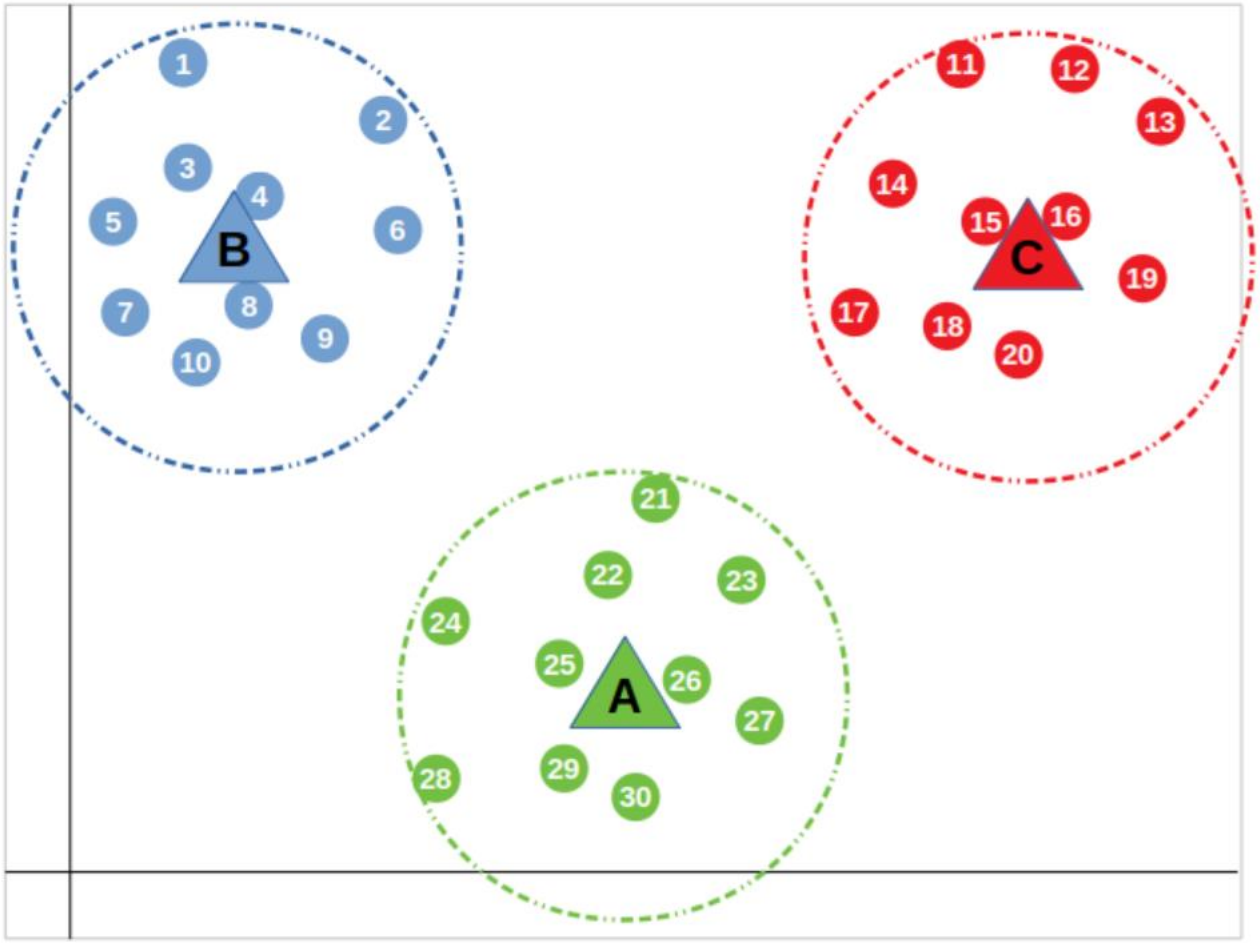

*Source: Anas Al-Masri*

# K-means or PAM




*Source: Anas Al-Masri*

# K-means or PAM




*Source: Anas Al-Masri*

# Step 3: Clustering, cont.

Diagnostics:

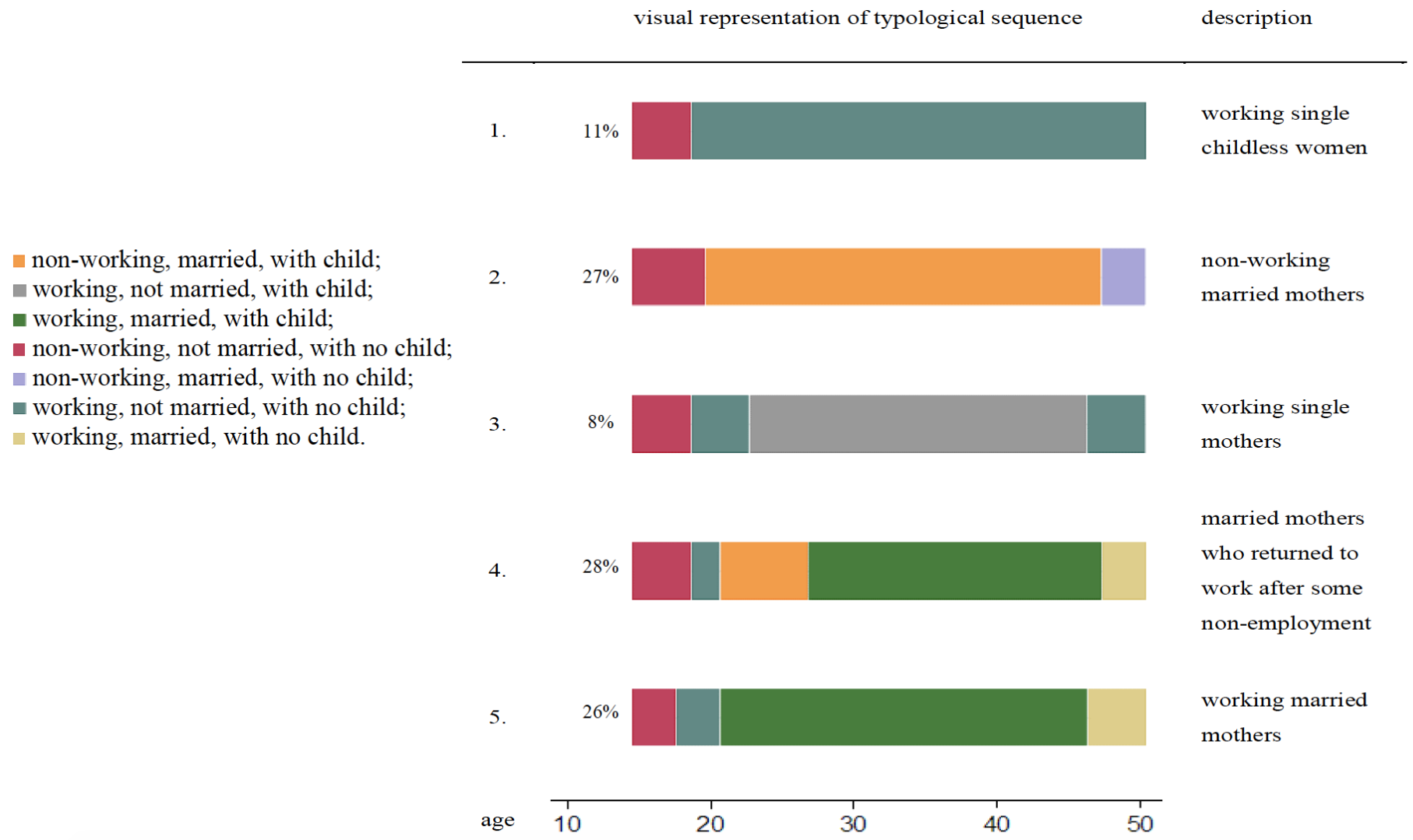
Goodness of fit tests

- Average silhouette width,
- Calinsky-Harabasz pseudo F-test

For more:  Kaufmann and Rousseeuw, 2009

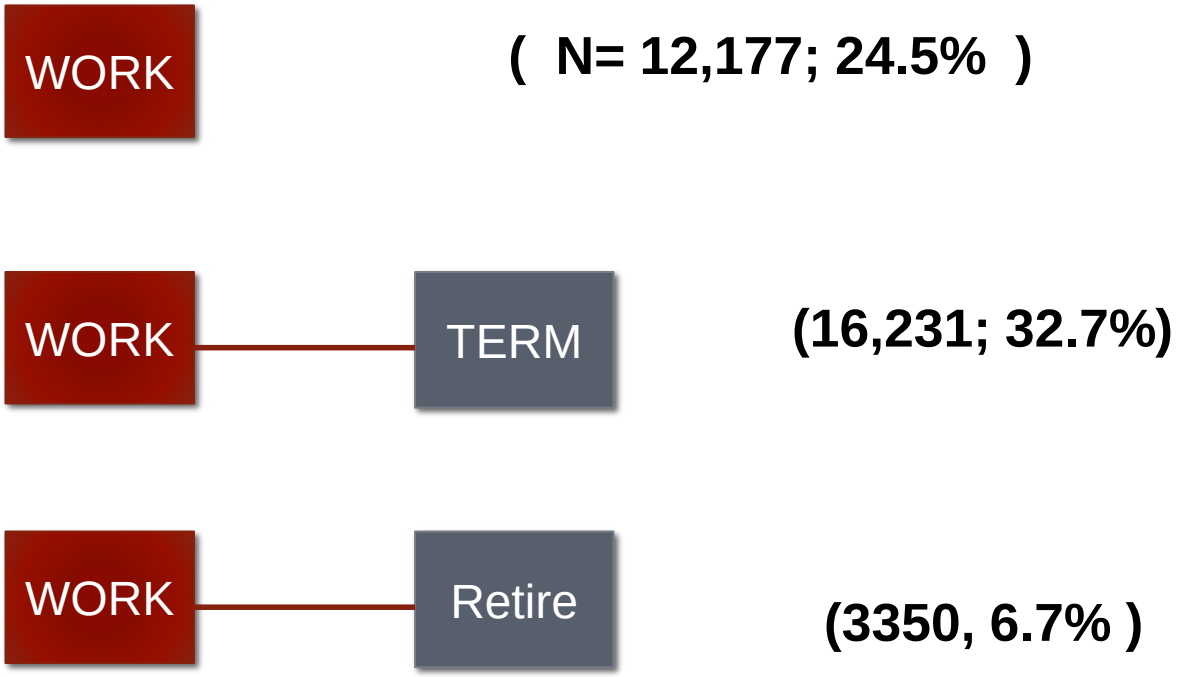
# Graphic representations for sequences and clusters

# Applications to Work-Life Trajectories



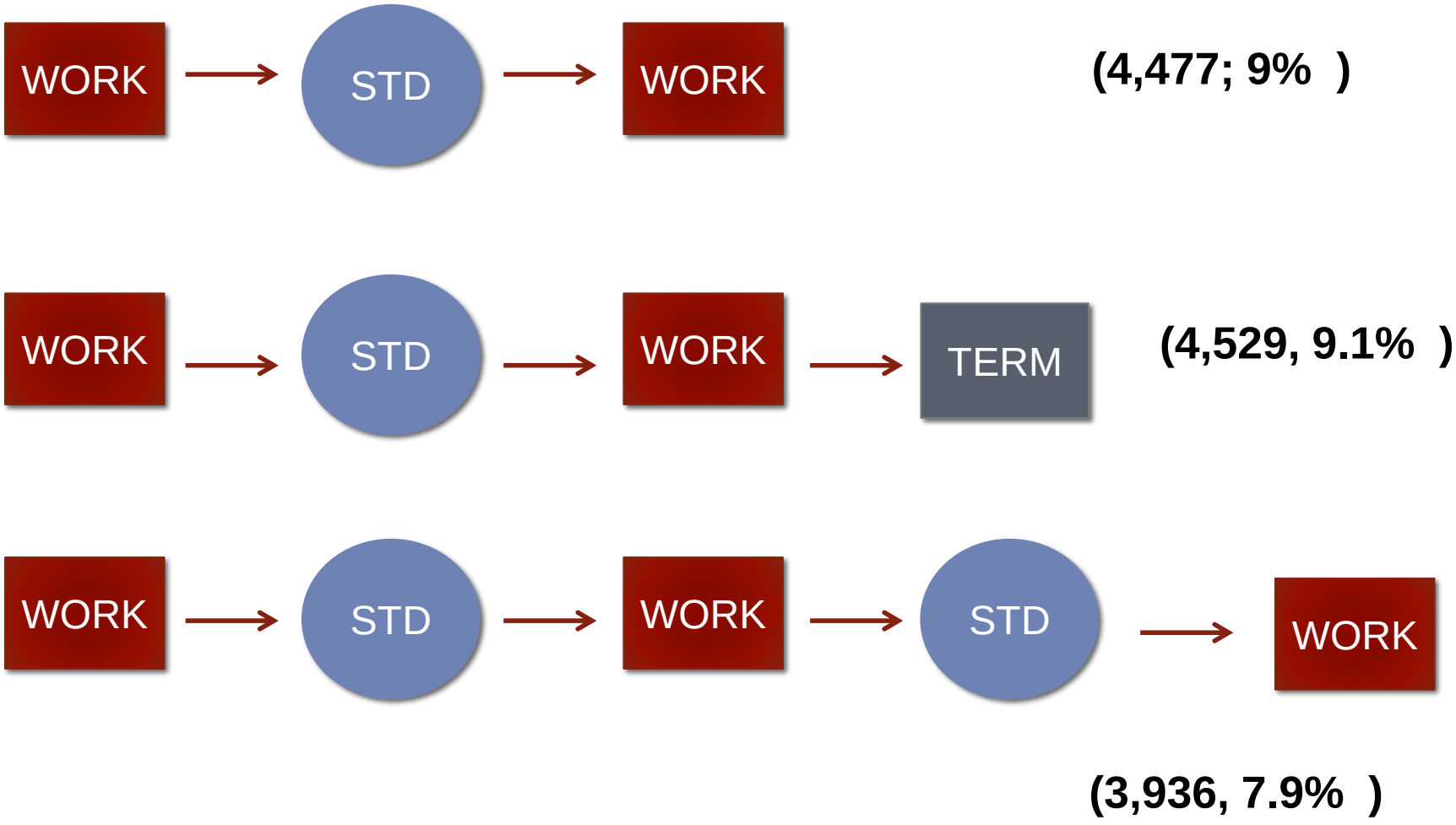

**Source:** van Hedel, Mejía-Guevara, et al. (Revise and resubmit – *American Journal of Public Health*)

# Identifying typologies: “Continuous Work”




64% (N=31,758) of the sample stays continuously employed (and/or leave Alcoa for other reasons) through sample period.

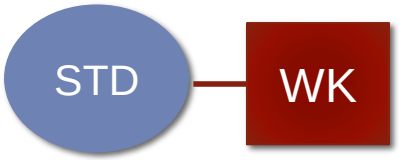
# Identifying typologies: “Minor STD”



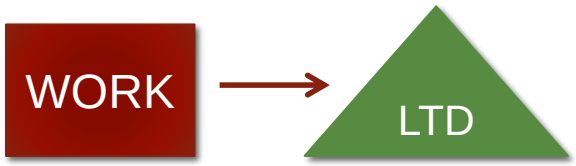
# Identifying typologies: “Disruptive”


**( 2,931; 5.9% )**

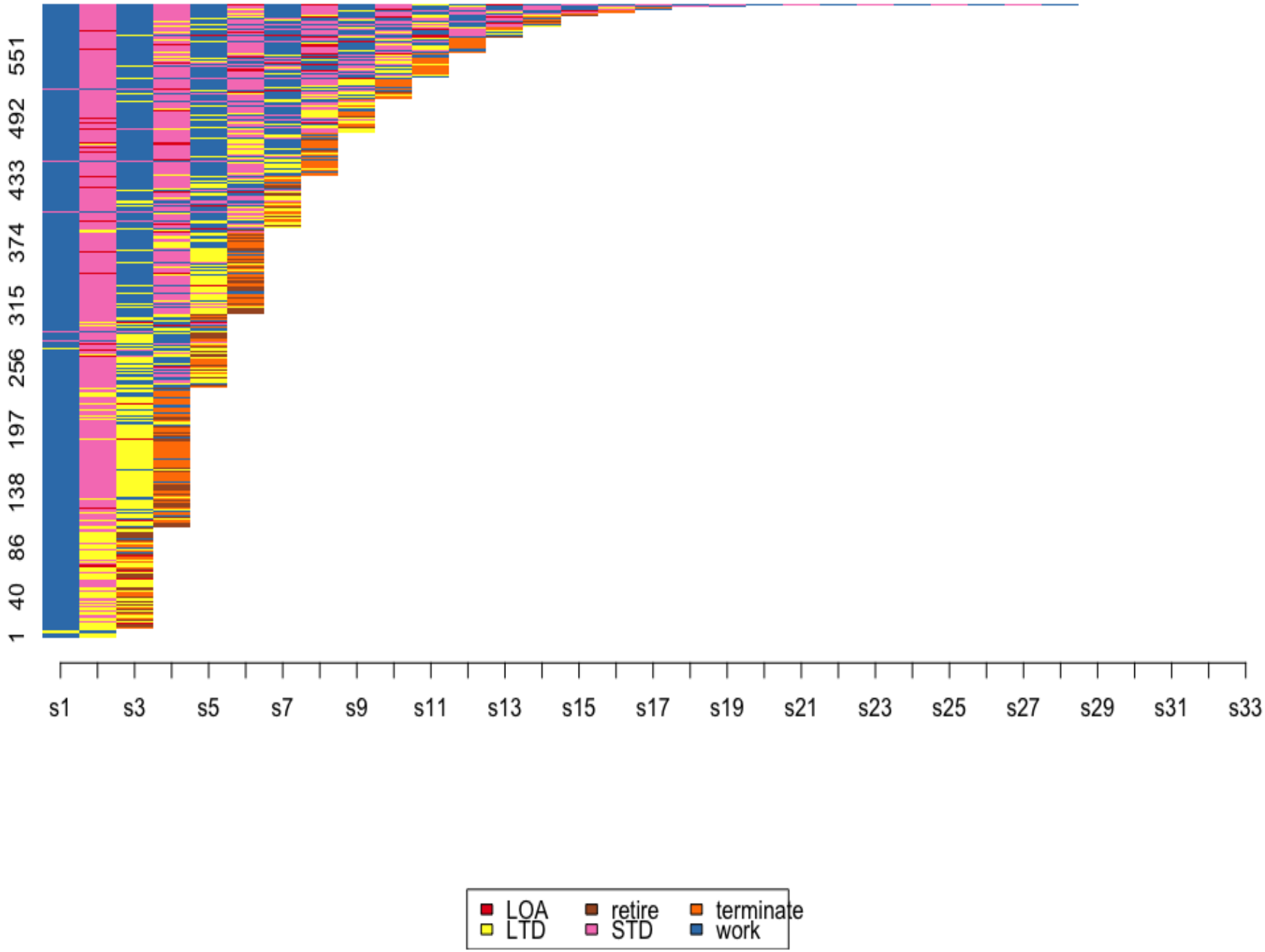

**( 1,262; 2.5% )**

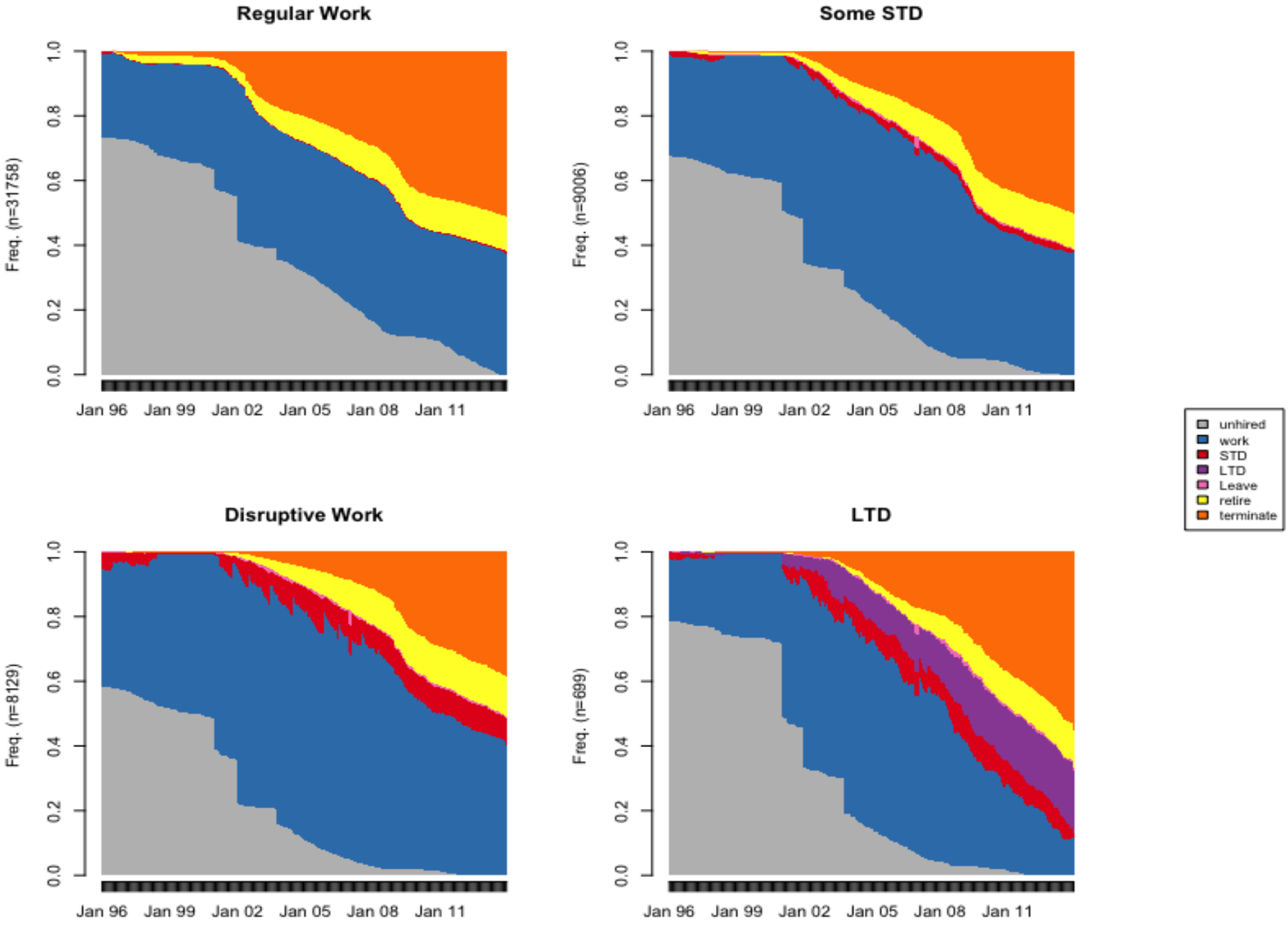


# Identifying typologies: “Ever LTD”




**( 1.4% )**


1
40
86
138
197
256
315
374
433
492
551
s1
s3
s5
s7
s9
s11
s13
s15
s17
s19
s21
s23
s25
s27
s29
s31
s33
LOA
LTD
retire
STD
terminate
work


Regular Work
Some STD
Disruptive Work
LTD
Freq. (n=31758)
Freq. (n=9006)
Freq. (n=8129)
Freq. (n=699)
0.0
0.2
0.4
0.6
0.8
1.0
Jan 96
Jan 99
Jan 02
Jan 05
Jan 08
Jan 11
unhired
work
STD
LTD
Leave
retire
terminate

# Reflections on the readings, QA, discussion

# Thank you!

Questions?

aharrati@mathematica-mpr.com