Dealing with missing states. A comparison of alternative procedures.

Adrien Remund
adrien.remund@unige.ch

Issue

- Many panel datasets are released on calendar time.
- Life course theories assume age-related critical periods (e.g. transition to adulthood).
- Age alignment is therefore crucial, but generates many missing states.

Research question

What are the advantages and drawbacks of different procedures to deal with missing data created by age alignment?

Data & Methods

- Swiss Household Panel, waves 1-12. Self-reported health (SRH): P$$C01. Age alignment: 15-30.
- 5'290 sequences with at least one valid observation. 79.3% missing, of which 56.7% 'abs' (not asked) and 22.6% 'na' (not responded).
- R-libraries : `TraMineR 1.8` *(Gabadinho et al. 2011)*, `WeightedCluster 0.9` *(Studer 2012)*, and `mice 2.12` *(van Buuren et al. 2011)*.
- Optimal Matching with custom substitution costs and comparison of the resulting distances and cluster solutions (Ward).

Results

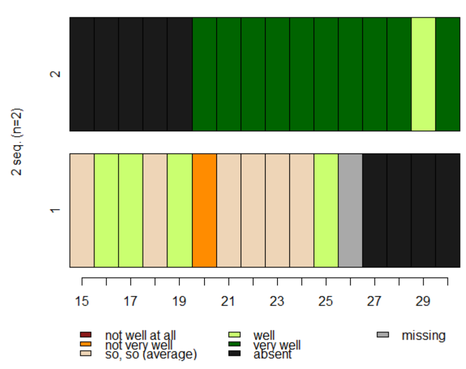
1. Usual treatment

Substitution costs matrix (indel = 2)*

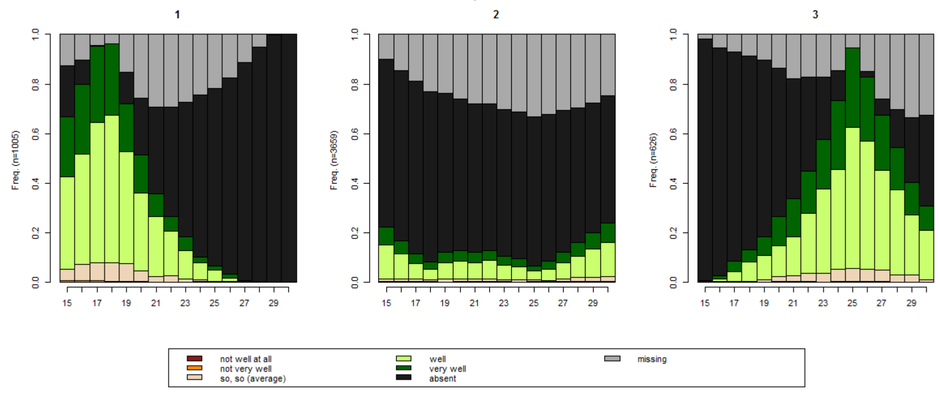
	very well	well	so,so	not very well	not well at all	na/ abs
very well	0.0	0.6	1.5	2.9	4.0	2.0
well	0.6	0.0	0.9	2.3	3.4	2.0
so, so	1.5	0.9	0.0	1.4	2.5	2.0
not very well	2.9	2.3	1.4	0.0	1.1	2.0
not well at all	4.0	3.4	2.5	1.1	0.0	2.0
na/abs	2.0	2.0	2.0	2.0	2.0	0.0

**Distances based on (Perneger et al. submitted for publication).*

Most different sequences



*Cluster representation**



* The 3 cluster solution locally maximizes the ASW, PBC and HC criterions.

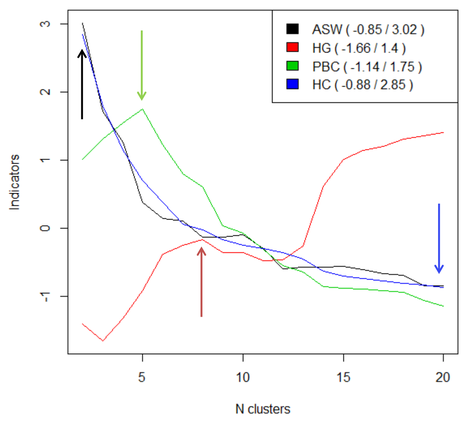
Conclusions: The distances and clusters are explained by the location of missing values.

2. Missing states as "jokers"

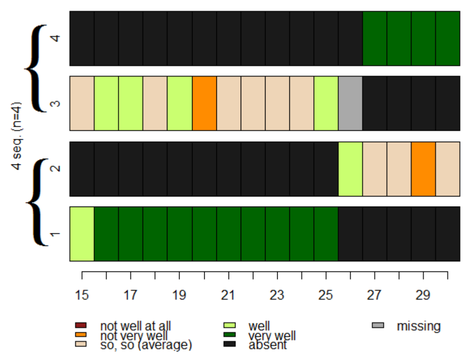
Substitution costs (SC) (indel = 2)

$$SC_{abs/na \rightarrow i} = SC_{i \rightarrow abs/na} = 0 \ \forall \ i$$

Cluster optimization



Some sequences with null distances



Conclusions: The clusters are independent of the na's location, but they are inconsistent (impossible to find an adequate number of clusters). Some distances that were computed as null, due to the location of the missing states, seem counterintuitive. Moreover, the null substitution costs break the triangle inequality.

3. Imputation

Predictors of state at time t (S_t)

$$S_t \sim \sum_{k=1}^{t-1} S_k$$

Parameters

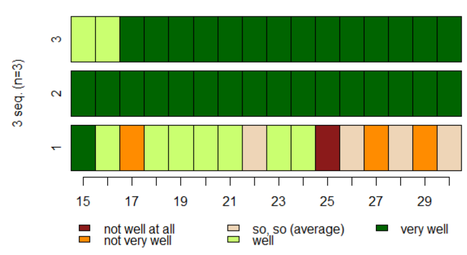
Predictive Mean Matching
Visit sequence: chronological
5 iterations
1 imputation

Distribution of SRH (%, excl. na's)

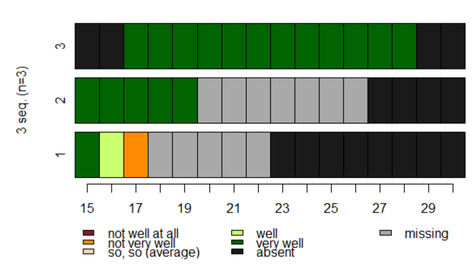
	original dataset	imputed dataset
very well	33.0	32.3
well	58.5	58.1
so, so	7.8	8.9
not very well	0.7	0.7
not well at all	0.1	0.1

Most different sequences

Imputed data

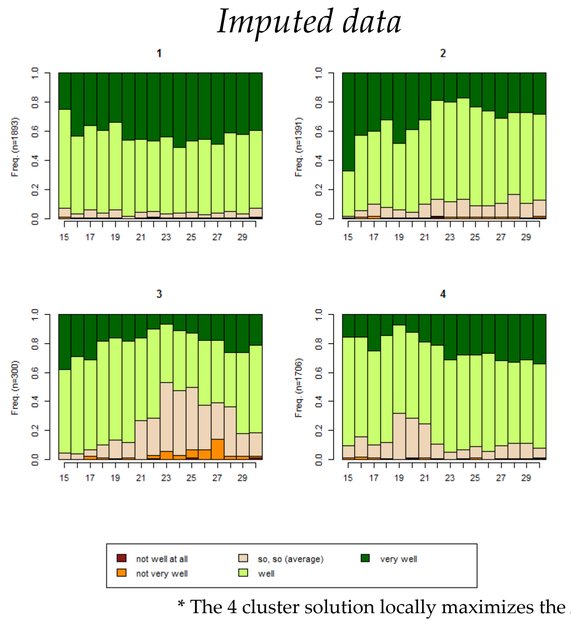


Original data

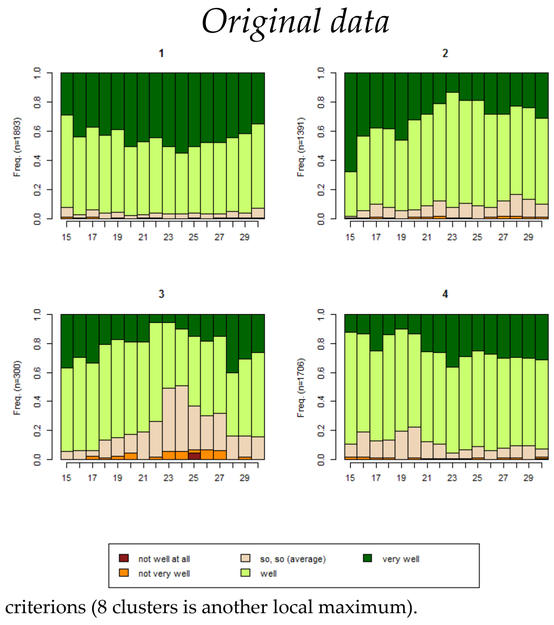


*Cluster representation**

Imputed data



Original data



* The 4 cluster solution locally maximizes the ASW, PBC and HC criterions (8 clusters is another local maximum).

Conclusions: At the micro level, imputation draws somehow arbitrary sequences, but the big picture makes much more sense. Since the representation of the original data (based on the new clusters) is almost identical to the imputed data, the clusters depict existing patterns.

References

Gabadinho, A., G. Ritschard, N. S. Müller and M. Studer (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 40(4).
Van Buuren, S., Groothuis-Oudshoorn, K. (2011). `mice`: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3).
Studer, Matthias (2012). *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles*, Thèse de doctorat n°777, Faculté des sciences économiques et sociales, Université de Genève.
Perneger, V., C. Combescure, A. Gayet-Ageron, D. Courvoisier, T. Agoritsas, S. Cullati (submitted for publication), *Rating one's health as excellent to poor: analysis of distances and transitions between response options*.