Imputing Sequence Data

Brendan Halpin, University of Limerick

Lausanne Conference on Sequence Analysis, 8 June 2012

Outline

Outline

Lifecourse data is gappy

- Longitudinal data is
  - gap prone: bias and loss of sample size
  - full of good info to impute from
- This talk presents an approach to imputation of life course data
  - that takes order into account
    - respects the longitudinality
    - exploits the longitudinality
  - that is useful to view in a multiple-imputation context

Goals of the talk

- Outline the imputation model
- Propose a way to use multiple imputation in a sequence-analysis/cluster-analysis (SACA) context
- Briefly compare MI with other ways of dealing with missing data

Outline

The idea of multiple imputation

- Standard approaches to missing data are inadequate
  - complete case analysis introduces bias and discards information
  - Mean imputation retains information but is also biased
- Regression-based imputation with a good predictive model is better but understates variability
- Rubin (1987) proposed "multiple imputations" drawn from the predictive probability distribution

"Rubin's Rules"

- Fit models on each imputed data set and average the parameter estimates
  - variance is average variance plus the variance between the estimate
- Imputed data serves as a non-biasing placeholder to allow full use of observed data

Why it doesn't work out of the box for SA

- Many variables to impute
  - E.g., 5 years of monthly data $\Longrightarrow$ 60 similar variables and up to 60 models to fit
- Most of the variables are highly collinear
- MI by "chained equations" would suggest one prediction equation per incomplete variable
- Standard approaches will not necessarily respect the longitudinal structure
  - E.g, `AAA....BBBB` may be imputed as `AAABABABBBB` whereas a single transition such as `AAAAABBBBBB` is more realistic

Longitudinally-aware recursive algorithm

- Treat data as a single variable, multiply observed
- Impute the gap, not the missing variable; but fill it in incrementally
- Control the order of the chaining to respect the longitudinal structure, closing the gap from its edges
- Key predictors are last and next observed states
  - The values of the state variable
  - Other measures relevant to those time points
- This approch targets SACA but should be valid for other analyses

Chained gap-healing

- Begin with longest gap, predict first (or last) element
- Then predict last (or first) of next shortest gap length (including longer gaps already reduced)
- Until no gaps remain

- Important to begin fill from edges
  - Least distance from observed data
  - But each gap has two edges: to begin pick one at random and impute
  - Then the other edge (of the newly shortened gap) has better data than the former, so alternate

- Only one predictive model per unit of longest gap

Sketching gap closure

- Five unit gap
- `XXX.....YYY`

Three unit gap

`XXX...YYYYY`

Sketching gap closure

- Five unit gap
- XXX.....YYY
- XXX....iYYY

Three unit gap

XXX...YYYYY

XXX...YYYYY

Sketching gap closure

- Five unit gap
- XXX.....YYY
- XXX....iYYY
- XXXi...IYYY

Three unit gap

XXX...YYYYY

XXX...YYYYY

XXX...YYYYY

Sketching gap closure

Five unit gap	Three unit gap
XXX.....YYY	XXX...YYYYY
XXX....iYYY	XXX...YYYYY
XXXi...IYYY	XXX...YYYYY
XXXI..iIYYY	XXX..iYYYYY

Sketching gap closure

- Five unit gap
- XXX.....YYY
- XXX....iYYY
- XXXi...IYYY
- XXXI..iIYYY
- XXXIi.IIYYY

Three unit gap

XXX...YYYYY

XXX...YYYYY

XXX...YYYYY

XXX..iYYYYY

XXXi.IYYYYY

Sketching gap closure

- Five unit gap
- XXX.....YYY
- XXX....iYYY
- XXXi...IYYY
- XXXI..iIYYY
- XXXIi.IIYYY
- XXXIIiIIYYY

Three unit gap

XXX...YYYYY

XXX...YYYYY

XXX...YYYYY

XXX..iYYYYY

XXXi.IYYYYY

XXXIiIYYYYY

Minimal model

- Multinomial logit

$$\log \frac{P(s_t = j)}{P(s_t = J)} = \alpha_j + \sum \beta_{1j}^k (s_{t-\delta_1} = k) + \sum \beta_{2j}^k (s_{t+\delta_2} = k)$$

- One of δ_1 and δ_2 is 1, the other the gap length
- The imputed value is drawn at random from the state space, following the predicted probability
- Previously imputed values are used to predict but not to estimate the model

Better models

- Next and last states are essential for longitudinal continuity and are powerful predictors
- But this base model makes several unrealistic assumptions
  - Time doesn't matter
  - Sequence history doesn't matter
  - Individual differences don't matter

Time matters

- Calendar time may matter: e.g., labour market transitions affected by state of economy
- Developmental time may matter: e.g., school-to-work trajectories marked by initial instability which declines
- Easily incorporated as a non-linear time effect: how observed transition rates change through time
- Could import external information about e.g., labour market conditions

History: non-Markov

- Spending time in a state earlier may make you more likely to stay in it or return to it
- Individuals acquire characteristics (history) that affect their transition rates
- By analogy, their future state distribution also affects the present
  - Not in a causal manner but prediction is concerned with joint distribution rather than causality
- Use before and after summaries: e.g., cumulative proportion of time in each of $J-1$ states

Individual heterogeneity

- Different types of individuals (gender, cohort, social class of origin) will have different transition matrices
- Take into account observed characteristics in the model
- E.g., add gender, class of origin or other fixed individual variables to prediction model
- May add relatively little information on top of what is already accounted for in the sequence data
- Desirable perhaps to take account of unobserved heterogeneity
  - E.g., individual random effects model: capture individual heterogeneity insofar as evidenced in the sequence
  - But computationally expensive (time) and unstable (many models, all must converge)

Data structure

- Gaps may depend in how the data are collected
- If we have information on this it should be included
- In the example below, information is available about whether the month was
  - the date of interview, or
  - reported explicitly as the start or end of a spell
- A gap immediately preceded by an interview is unlikely to start a new spell, unlike a gap following an explicit end
- This sort of information can improve the model fit substantially

Outline

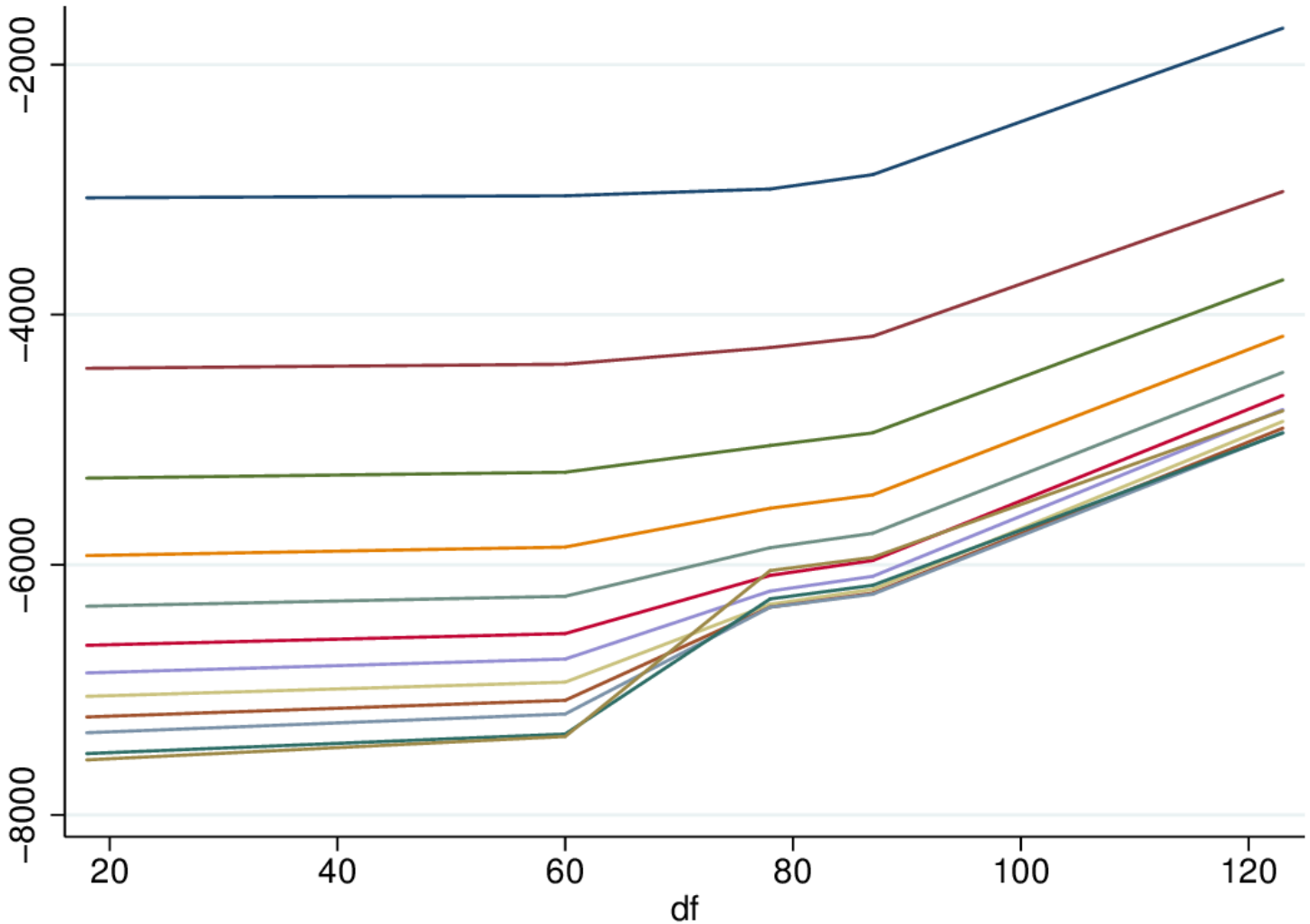
Example data set

- 6-year monthly labour market history of women who give birth at the end of year 2
- State space: Full-time, part-time, unemployed, non-employed
- Drawn from British Household Panel Study
- 1,096 are observed at the start and end of the period
- Of these, 157 have gaps, 149 of no more than 12 consecutive months
- Imputing up to 12 month gaps increases the sample from 939 to 1,088, nearly 16%

Models used

1. Prior and subsequent state
2. Prior and subsequent state interacted with sequence time (quadratic)
3. Plus prior and subsequent history
4. Plus data observation structure
5. Plus data observation structure interacted with prior and subsequent state

Model performance: log-likelihood



Model performance: predicted probabilities

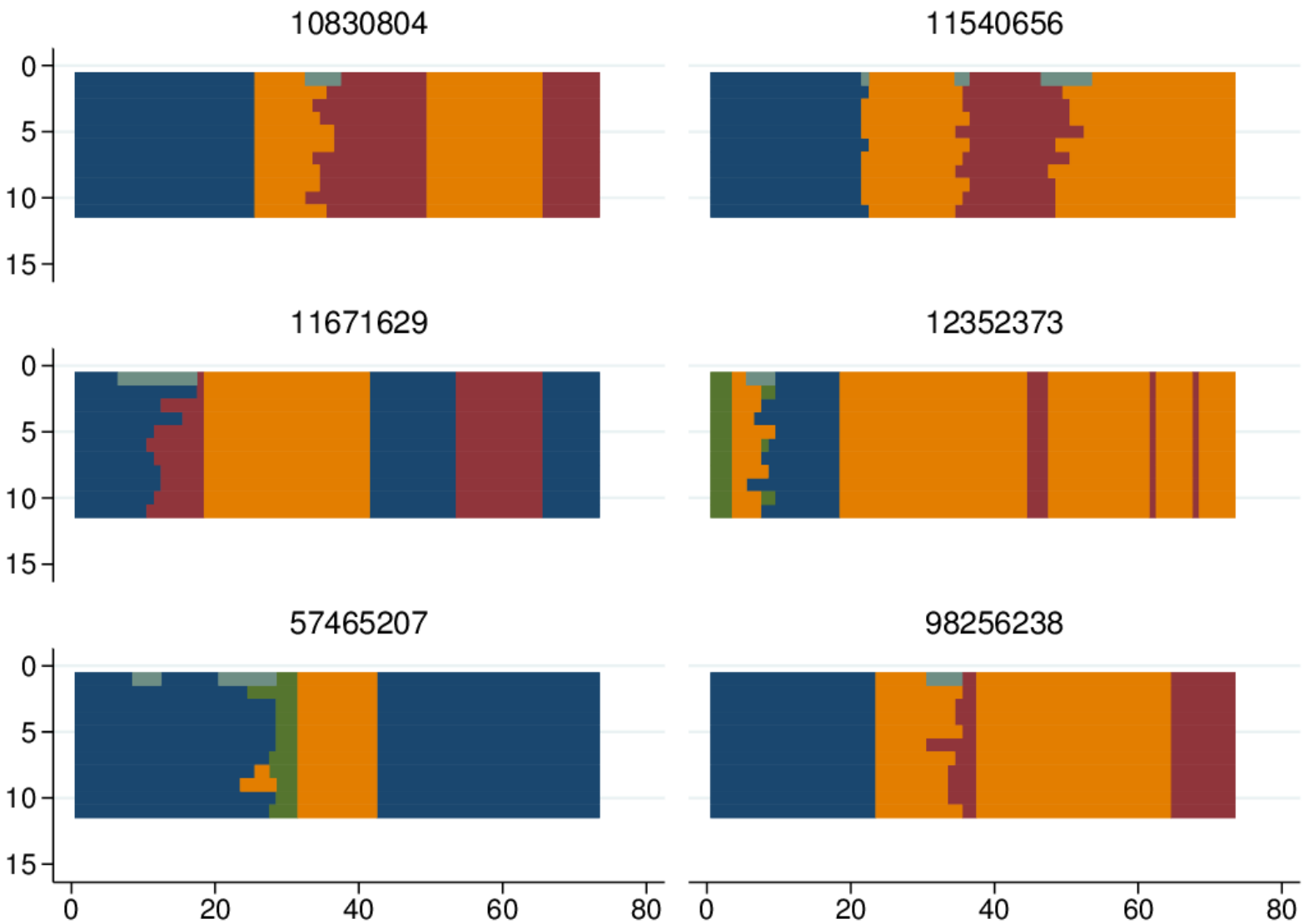
- The models differ only subtly in predicted probabilities
- Lowest correlation is in excess of 0.98, majority above 0.99
- Most predicted probabilities very close to 0.0 or 1.0
- Non-negligible minority in between
- Variation increases with gap length and complexity of model

Predicted probabilities

- For the most complex model
- And gaps of 1 and 12 units
- About 90-97% at extremes

Probability	Full-time	Part-time	Unemployed	Non-employed
Gap 1 unit				
0.00 - 0.01	64.04	80.46	95.90	54.90
0.01 - 0.98	2.27	2.19	0.92	2.85
0.98 - 1.00	33.69	17.34	3.18	42.25
Gap 12 units				
0.00 - 0.01	58.7	75.4	93.4	45.0
0.01 - 0.98	11.8	10.7	4.2	16.6
0.98 - 1.00	29.5	13.9	2.4	38.4

Some example imputations



Pattern of imputations

- Clear from inspection that the imputation behaves in general as expected
- Gaps bracketed by a single state usually filled in with that state
- Gaps bracketed by two states randomise the timing of the transition across the gap
- Some extra transitions or third states imputed, but few
- More likely with longer gaps

Outline

Multiple imputation in a non-stochastic context

- We proceed as usual by generation of pairwise distance using OM or another algorithm, preparatory to cluster analysis
- OM under TraMineR or my SADI Stata plug-in deals efficiently with duplicates
- But MI yields greatest benefit in stochastic modelling
- How do we proceed when the analysis is non-stochastic?

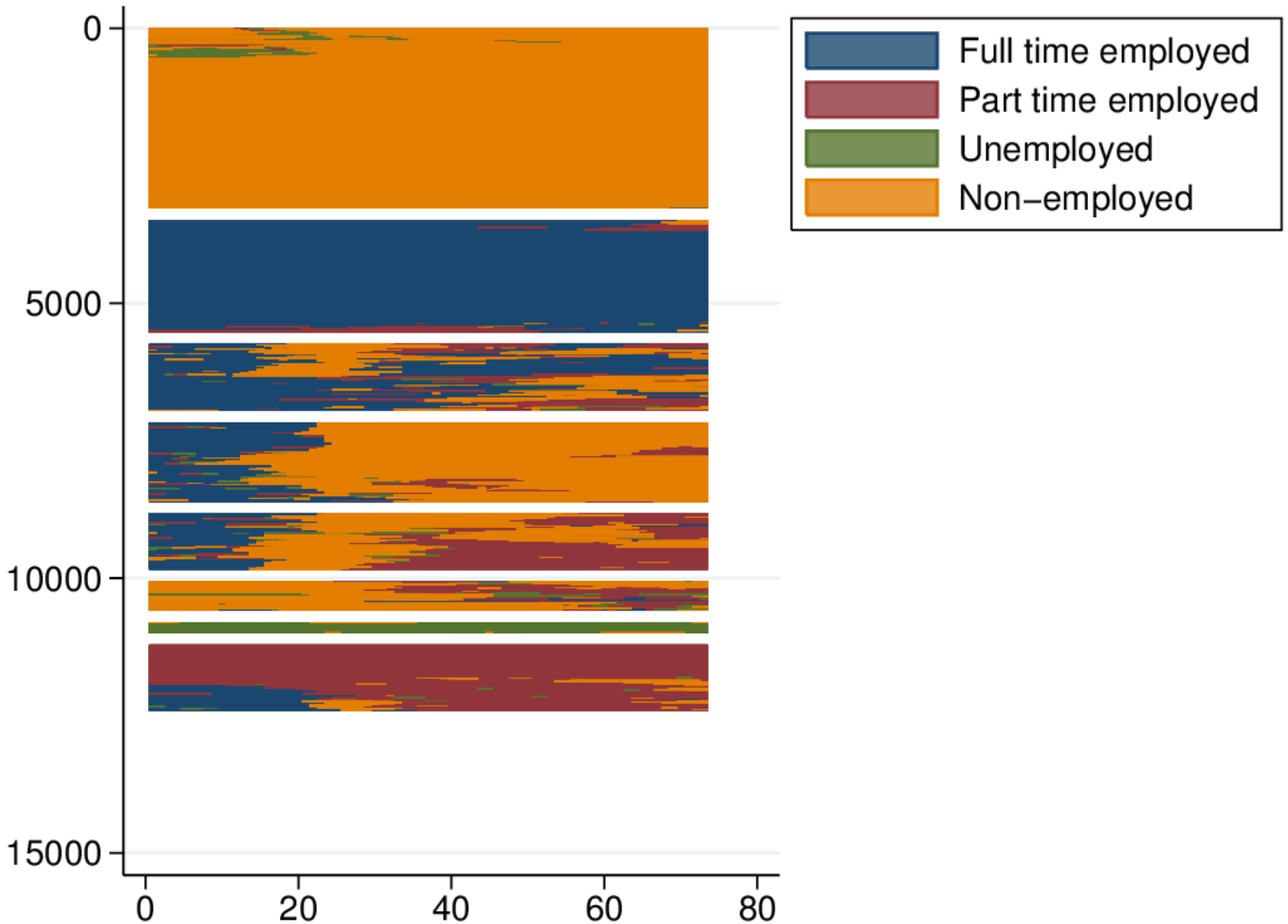
Clustering options

- Cluster the imputed data sets separately and compare: but cluster solutions across different data will be unstable (even small data differences)
- Pool the data sets and do a single big CA: slow, memory-intensive but more stable
- Average pairwise distances between imputations and other sequences – reduces to a single measure but hides variability altogether

Pooled cluster analysis preferred

- R replications means R copies of the full sequences and R instances of imputed ones
- Ward's clustering not distorted by multiple instances
- Achieve a measure of uncertainty
  - By looking at difference in cluster membership of imputed sequences
  - By looking at difference in distances between imputed sequence and all others
  - This leads naturally to the *discrepancy* measure (Studer et al, 2011); for future work
- In what follows I do the first of these

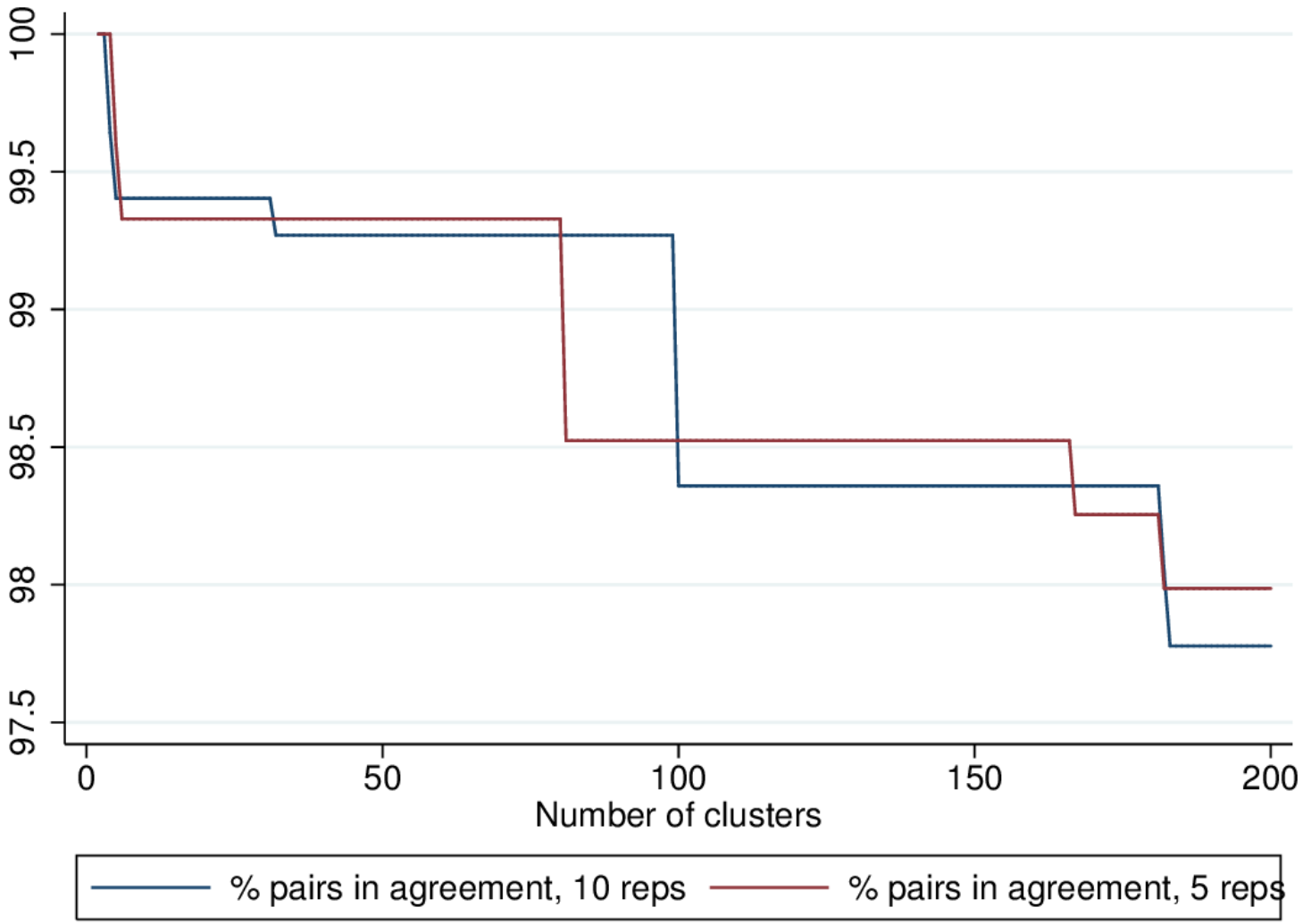
8-cluster solution, 10 replications



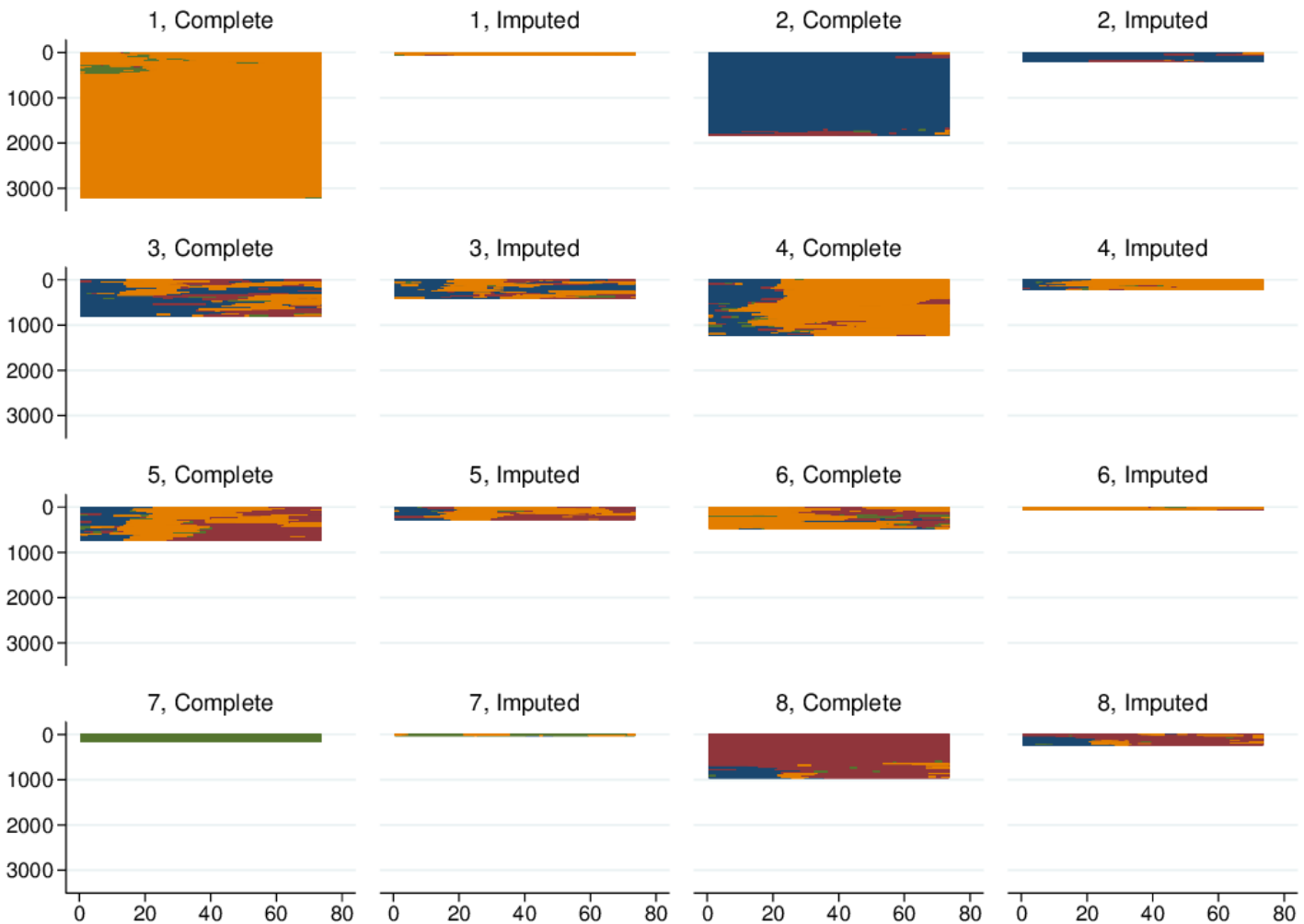
Variation in cluster membership

- For imputations of the same sequence, count every pair in the same cluster
- Even with 50 clusters, >99% of pairs are in the same cluster
- Partly because some imputations are identical
- Partly because imputations affect a small part of the sequence
- And tend to differ in ways that are discounted by OM
- Other data sets will generate different results, but the picture here is that gaps can be imputed with relatively inconsequential variability

Assessing variation



Complete vs imputed sequences



Complete vs imputed sequences

- Imputed sequences are distributed differently across the clusters
- Much less in the "stable" clusters
- More or much more in "interesting" clusters (3, 4, 5, and the "interesting" part of 8)
- Very likely that people with changing employment status are less likely to be present at each interview
- Also likely gaps where the employment status is stable are easier to cover in the interview
- Clear that not imputing distorts the sample, and reduces the number of "interesting" sequences to analyse

Two approaches to compare

- Use average distance across imputations
  - Better than assigning to most probably state (single imputation)
  - Generates a single manageable distance matrix
- Code for missing, with a maximal substitution cost for pairs involving missing
  - Including missing/missing
  - Resulting distance is a ceiling

Average distance

	Avg							
MI	1	2	3	4	5	6	7	8
1	2730	—	—	—	—	—	540	—
2	—	1940	100	—	—	—	—	—
3	—	—	850	—	220	—	—	140
4	—	—	—	850	254	—	330	—
5	—	—	—	130	884	—	—	—
6	—	—	—	—	—	530	—	—
7	—	—	—	—	—	190	—	—
8	—	—	—	—	2	—	—	1190

- κ_{max}: 0.79
- Adjusted Rand Index: 0.77

Missing as maximally different

	Avg							
MI	1	2	3	4	5	6	7	8
1	3250	—	—	20	—	—	—	—
2	—	2010	20	—	—	—	10	—
3	—	90	870	10	60	—	180	—
4	—	—	—	1400	24	10	—	—
5	—	—	50	90	854	—	10	10
6	30	—	—	80	30	390	—	—
7	—	—	—	—	—	190	—	—
8	—	—	—	—	32	—	450	710

- κ_{max}: 0.84
- Adjusted Rand Index: 0.88

Alternatives: tentative conclusions

- Results are similar but markedly not identical
- Average distance results are volatile:
  - Comparisons with separate groups of 5 replications vary quite a bit
- In this exercise the missing-as-different approach works well

Other directions

- Discrepancy
  - Avoids a lot of the discomfort of CA
  - Can we re-write Rubin's Rules for discrepancy?
- Better predictive models
  - Incorporate more individual or temporal information, individual random effects
- Extending at start and end
  - No reason not to fill short cantilever gaps at either end

Multiple domains: imputing across domains

- If information on one domain is available while another is missing
- If information on all domains is missing simultaneously
- Requirement that domain-specific longitudinality is protected at the same time as cross-domain coherence
- But potentially much better imputation: more information

Outline

Conclusion 1/2

- Dealing with gaps in sequence data is essential: lots of "interesting" sequences have gaps, ignoring them leads to bias and inefficiency
- The formal structure of the imputation works well, though attention needs to be paid to the predictive adequacy of the imputation model
- It may be sufficient to base the imputation on information in the sequences alone; individual-level data may be additionally useful but is not necessary

Conclusion 2/2

- Clustering pooled imputation data sets is feasible and gives interesting results, but is not as efficient as it might be
- Variation in cluster membership of replicated imputations is a useful measure of imputation uncertainty
- Other alternatives that hide the uncertainty are probably less attractive