Missingness in Sequence Analysis

Brendan Halpin, University of Limerick

Algorithmic Methods in Social Research
Nuffield College, Oxford, 27 Feb 2015

http://teaching.sociology.ul.ie/seqanal/nsi_missing.pdf

The problem of missingness in lifecourse data

- Longitudinal data is very susceptible to missingness
  - Duration models cope well with right-censoring, but there is still information loss
  - Holistic approaches absolutely need complete histories
- I address two types of missingness
  - missingness *per se* (gaps)
  - sequence truncation (late entry or early exit)
- I propose alternatives based on coding missing (giving missingness a location in the state space), and introduce the notion of a "non-self-identical" missing value

Existing solutions

- Treating missingness as a special state has already been suggested (e.g., TraMineR manual) but how do we set its values?
- For gaps, the copy-value-forward strategy is often used but is not always justifiable
- For gaps, multiple imputation works but is onerous
- For truncated sequences, OM can deal with them automatically but is likely to sort short with short and long with long sequences, independently of their substance
- New proposed solution:
  - Locate missing as a neutral state in the state space
  - The notion of "non-self-identical" missingness

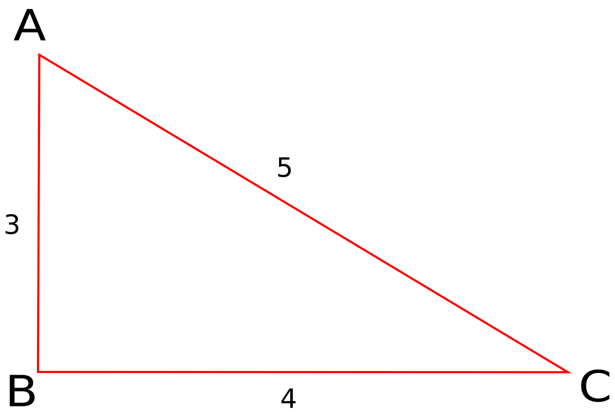
Missingness: bounded uncertainty?

- Missing is an unknown state
- Sometimes this doesn't matter
  - (TRUE or MISSING) resolves to TRUE
  - (FALSE and MISSING) resolves to FALSE, etc.
- In quantitative contexts, we can sometimes give upper- and lower-bounds to calculations involving missingness

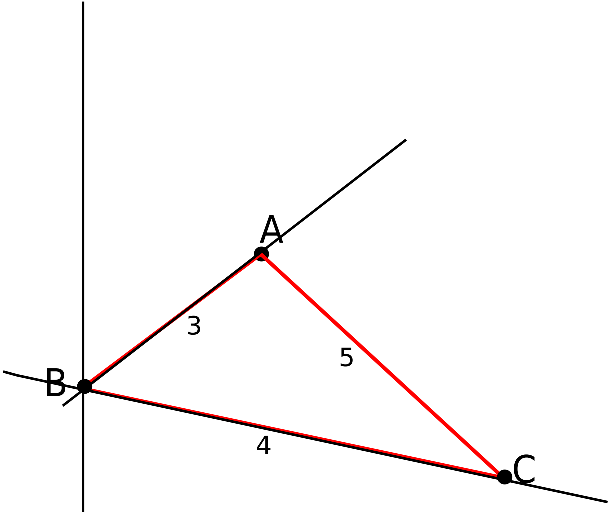
Treat missing as a point in the space, but an odd one

- We desire to give missingness a neutral effect
  - Treat it as a category that is equidistant from all others
- Maximally distant?
  - Penalise sequences for containing missings: estimated distance is at least as much as the true value and probably more
- Minimally distant?
  - This pushes missingness to disappear: estimated distance is as low as possible, probably lower than the true value
- Let's think geometrically, as if (for now) we had Euclidean distances

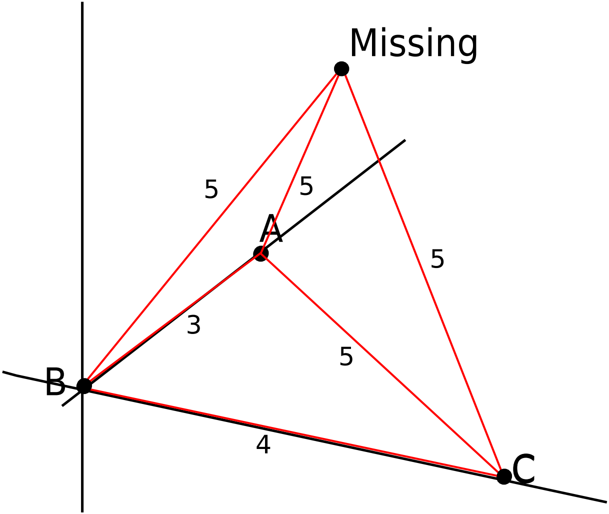
A 3-state space, in 2D



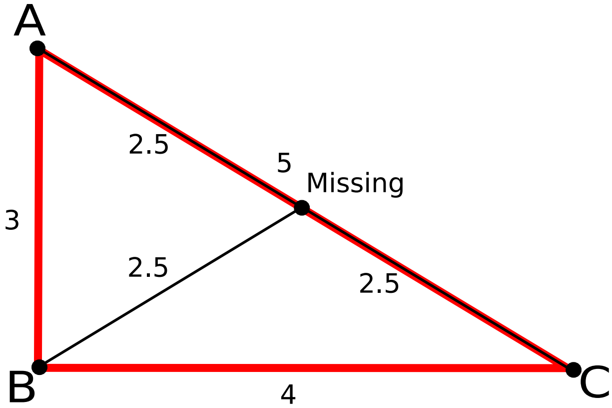
A 3-state space, in 3D



Missing as maximally distant



Missing as minimally different (Euclidean)



Substitution matrix, maximally different

	A	B	C	?
A	0	3	5	5
B	3	0	4	5
C	5	4	0	5
?	5	5	5	0

Substitution matrix, minimally different

	A	B	C	?
A	0.0	3.0	5.0	2.5
B	3.0	0.0	4.0	2.5
C	5.0	4.0	0.0	2.5
?	2.5	2.5	2.5	0.0

Substitution matrix, state-specific maxima

	A	B	C	?
A	0	3	5	5
B	3	0	4	4
C	5	4	0	5
?	5	4	5	0

Substitution matrix, state-specific minima

No longer Euclidean but still a distance:

	A	B	C	?
A	0.0	3.0	5.0	2.5
B	3.0	0.0	4.0	2.0
C	5.0	4.0	0.0	2.5
?	2.5	2.0	2.5	0.0

Metric spaces and missingness

- We have a natural upper bound for d(X,?) as the maximum distance
- It is possible to exceed it but will penalise missingness excessively
- The lower bound cannot be the minimum distance (which is zero)
  - Rather, the triangle inequality demands half the maximum
- Logically, it seems attractive to use state-specific maxima

Non-self-identical missing

- If we treat missing as a separate category, by default we consider missing a match with missing: d(?,?) = 0
- This is logically incorrect: missing should be non-self-identical:

	A	B	C	?
A	0.0	3.0	5.0	2.5
B	3.0	0.0	4.0	2.0
C	5.0	4.0	0.0	2.5
?	2.5	2.0	2.5	2.5

NSI: non-metric

- To treat missing as NSI implies a non-metric substitution cost matrix:
  - A non-zero on the diagonal: d(X,X) should always be zero
- But clearly d(?,?) is not necessarily a distance between identical states
- Define the result as an estimate of a metric distance, with measurement error

Simulations

- I now proceed by running a series of simulations
- Test varieties of "coded missing" against existing approaches
- For truncation and gaps
- Patterns of missing are imposed at random on observed data
- Compare several solutions in terms of how close they are to the full data

Simulating delayed entry

- Two data sources
  - Mothers' labour market history (BHPS, 72 months, 4 states, birth at end of year 2)
  - McVicar/Anyadike-Danes transition from school to work data (Northern Ireland, 72 months, 6 states)
- Simulation over-writes the first 0-19 months with a pre entry state, at random
- Thus no association between pre-entry duration and substance of sequences
- Compare the results on a number of measures
  - Measure of cluster agreement: Adjusted Rand Index
  - Correlation between inter-sequence distances
  - ANOVA testing association between duration of pre-entry and 8- and 16-cluster solutions: significant association is not desired

Mothers' labour market history, moderate missing

Average scores across 1,000 simulations:

	ARI	Corr	ANOVA-8	ANOVA-16
Full data	1.000	1.000	0.806	0.899
Censored data	0.721	0.950	0.804	0.901
Unequal lengths	0.679	0.985	0.184	0.000
SI, min	0.521	0.973	0.000	0.000
NSI, min	0.546	0.974	0.000	0.000
SI, max	0.552	0.975	0.000	0.000
NSI, max	0.658	0.985	0.113	0.000
NSI, state-specific max	0.598	0.982	0.000	0.000
NSI, state-specific min	0.529	0.974	0.000	0.000
SI, state-specific min	0.542	0.985	0.000	0.000
SI, state-specific max	0.596	0.982	0.000	0.000

Mothers' labour market history: best measure per simulation

	ARI	Corr	ANOVA-8
Censored data	562	0	0
Unequal lengths	235	934	531
SI, min	0	0	0
NSI, min	1	0	8
SI, max	1	0	8
NSI, max	127	0	296
NSI, state-specific max	2	0	37
NSI, state-specific min	0	0	0
SI, state-specific min	1	0	21
SI, state-specific max	6	1	34

MVAD data, moderate missing

	ARI	Corr	ANOVA-8	ANOVA-16
Full data	1.000	1.000	0.806	0.902
Censored data	0.360	0.942	0.808	0.906
Unequal lengths	0.622	0.982	0.642	0.321
SI, min	0.415	0.938	0.000	0.000
NSI, min	0.571	0.934	0.106	0.000
SI, max	0.589	0.944	0.175	0.001
NSI, max	0.609	0.982	0.674	0.116
NSI, state-specific max	0.564	0.977	0.131	0.000
NSI, state-specific min	0.591	0.978	0.195	0.000
SI, state-specific min	0.537	0.977	0.066	0.000

MVAD data: best measure per simulation

	ARI	Corr	ANOVA-8
Censored data	0	0	0
Unequal lengths	297	418	389
SI, min	0	0	0
NSI, min	97	0	16
SI, max	132	0	43
NSI, max	208	582	475
NSI, state-specific max	77	0	25
NSI, state-specific min	157	0	41
SI, state-specific min	32	0	11

Findings

- OM with unequal-length sequences does much better than I expected
- Whether censoring is effective depends on the data: MVAD is very different from the mothers' data
- NSI with maximal distance does well
- NSI does rather better than SI: the pre-entry state is not treated as similar to itself

Simulating general missingness

- We again use the mothers' labour market history data, with various levels of missingness
- Each month as P(missing) between 0.01 and 0.1, rising to between 0.3 and 0.7 if the previous month is missing
- This generates a realistic pattern of runs of missingness

Agreement, averages

	ARI	Corr
SI, min	0.803	0.990
NSI, min	0.808	0.990
SI, max	0.760	0.986
NSI, max	0.771	0.983
NSI, state-specific max	0.770	0.984

ANOVA, averages

	8-cluster	16-cluster
SI, min	0.426	0.295
NSI, min	0.440	0.351
SI, max	0.302	0.078
NSI, max	0.364	0.070
NSI, state-specific max	0.331	0.061

Best agreement, by run

	ARI	Corr
SI, min	113	303
NSI, min	148	97
SI, max	35	0
NSI, max	49	0
NSI, state-specific max	55	0

Best ANOVA, by run

	8-cluster	16
SI, min	95	126
NSI, min	126	228
SI, max	45	14
NSI, max	65	16
NSI, state-specific max	69	16

Findings: general missing

- Minimum distance does well, both NSI and SI: high kappa, ARI and correlation.
- On ANOVA, NSI does well.
- SI does best on correlation, but as missingness rises NSI does better (more missing/missing comparisons)

Real applications

Two applications:

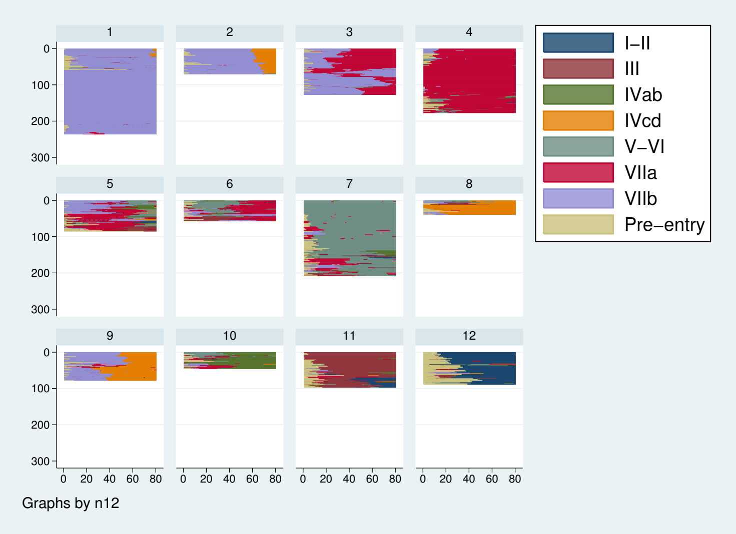
- Class career data (as in 1998 paper with Tak Wing Chan), with pre-entry. Irish males, class careers from 15 to 35 (1973 data)
- Mothers' labour market data, comparing with multiple imputation

Irish Mobility Study pre-entry

- Collected 1973, retrospective life histories, males only
- Coded in EGP class scheme

In 1998 paper we treat pre-entry as maximally similar to all other states, but excessively so, yielding non-metric distances.

Mid 20th century Irish class careers



ARI

Adjusted Rand Index comparing several 8-cluster solutions

	ESR1	ESR0	NSI min	NSIv	Unequal
ESR (1)	1.000				
ESR (0)	0.700	1.000			
NSI minimum	0.622	0.602	1.000		
NSI state-specific min	0.802	0.644	0.596	1.000	
Unequal lengths (OM)	0.887	0.724	0.595	0.762	1.000

Correlation

Correlation between distance matrices

	ESR1	ESR0	NSI min	NSIv	Unequal
ESR (1)	1.000				
ESR (0)	0.983	1.000			
NSI minimum	0.992	0.957	1.000		
NSI state-specific min	0.994	0.963	0.999	1.000	
Unequal lengths (OM)	0.990	0.959	0.997	0.996	1.000

Conclusion re IMS data

- NSI attractive but no gold standard to compare it with
- Unequal lengths OM yields results quite like non-metric 1998 results

Mothers' labour market history

- This data is very subject to general missingness: gaps
- Missing here a problem of
  - 1: losing cases
  - 2: bias: losing interesting cases.
- Here we have a gold standard: multiple imputation

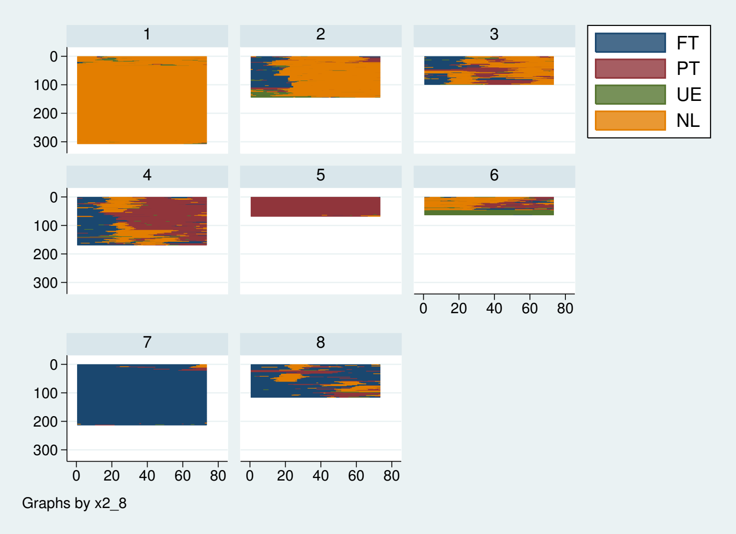
Multiple imputation: an aside

- Sequence data contains lots of information from which gaps can be imputed, often with little variability
- Collateral data (time-varying states in other domains, fixed characteristics) can improve the imputation
- But often the sequence itself contains enough data
- For more information see:
    - 'Multiple Imputation for Life-Course Sequence Data', 2012, `http://www.ul.ie/sociology/pubs/wp2012-01.pdf`
    - 'Imputing Sequence Data: Extensions to initial and terminal gaps, Stata's `mi`', 2013, `http://www.ul.ie/sociology/pubs/wp2013-01.pdf`

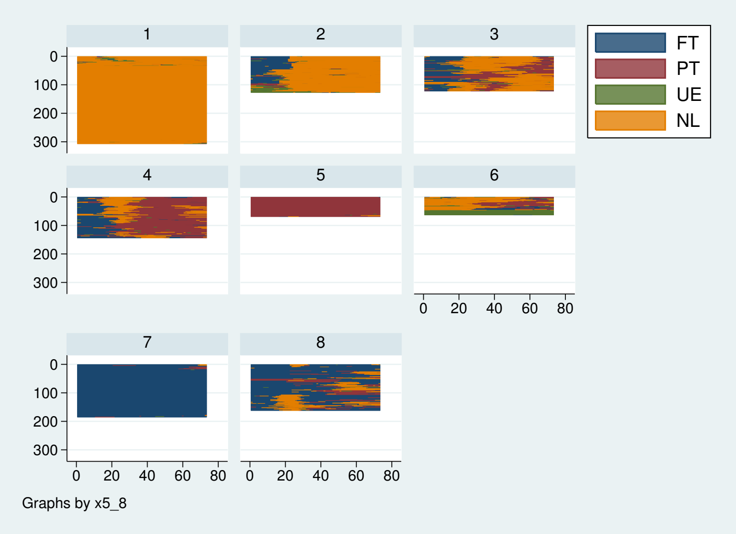
Basic distance matrix:

	FT	PT	UE	NL
FT	0	1	2	3
PT	1	0	1	2
UE	2	1	0	1
NL	3	2	1	0

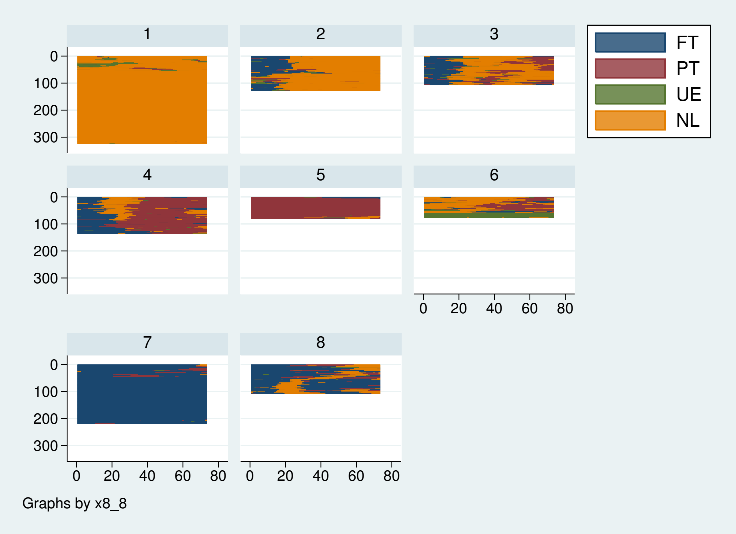
Imputed data, 8 cluster solution



Coded missing, NSI max



Coded missing, NSI var min



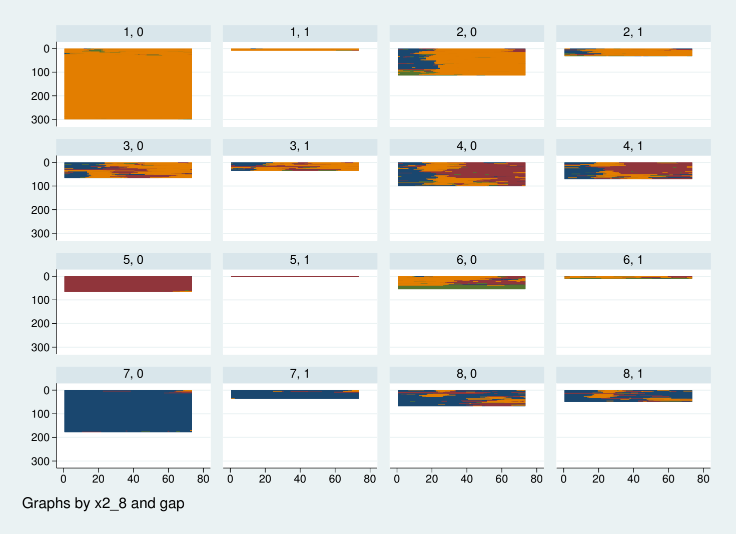
Coded missing vs imputed: correlation of distances

	Kappa	ARI	Correlation
Imputed	1.000	1.000	1.0000
SI min	0.776	0.715	0.9967
NSI min	0.729	0.720	0.9966
SI max	0.857	0.820	0.9909
NSI max	0.822	0.794	0.9903
NSI var max	0.762	0.754	0.9917
NSI var min	0.785	0.748	0.9968
SI var min	0.741	0.742	0.9969

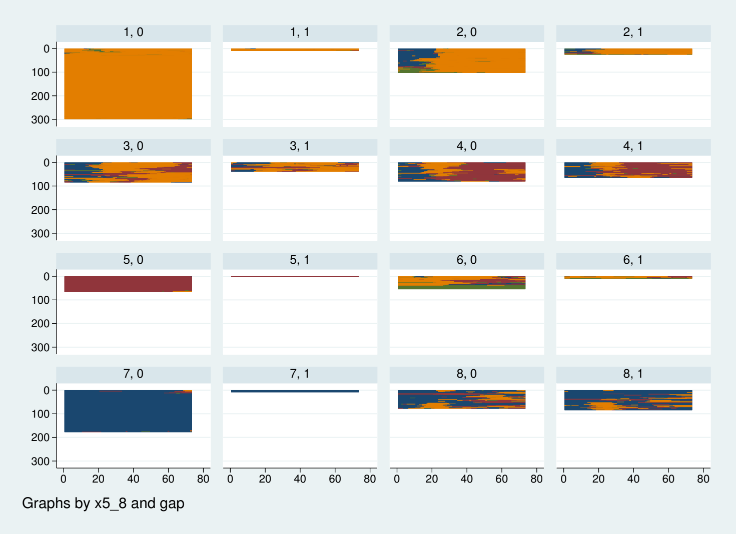
Gappy sequences are more complex

- Missingness is associated with cluster membership via its association between transition rates
    - volatile careers are more subject to missingness
- Ideally the method shouldn't add to this
    - by making missingness too different from everything else
    - by making missingness too self-similar

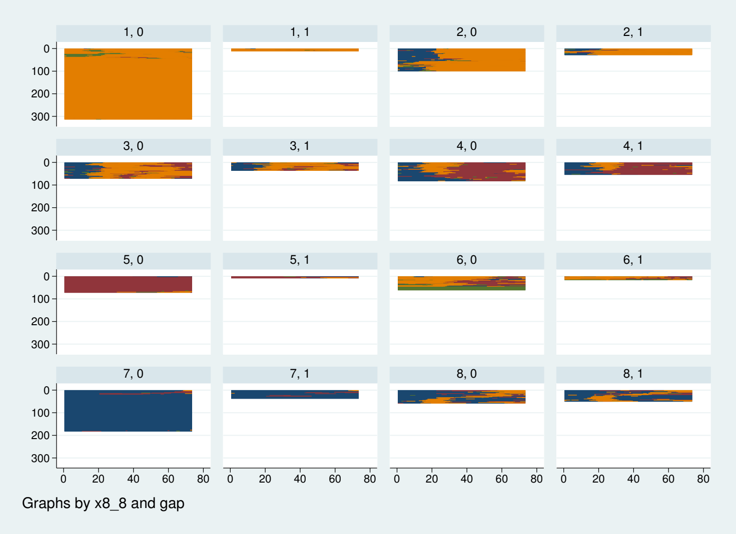
Imputed data, showing imputed vs whole sequences



NSI max: missing and non-missing



NSI var min: missing and non-missing



Association between gappiness and cluster

Method	χ^2
MI (single)	165.9
SI min	155.0
NSI min	125.7
SI max	258.6
NSI max	139.1
NSI v max	155.5
NSI v min	152.5
SI v min	117.4

Case Conclusion

- SI gives results close to imputation
- NSI max also good
- However, SI max shows signs of exaggerating the similarity of gappy sequences

Conclusion

- Treating missing as a special state is workable, but not a magic bullet
- NSI missing works well in some situations
- For truncation
  - Censoring is bad, at least with some forms of data
  - OM variable length comparison is surprisingly effective
  - In general NSI serves to make pre-entry more neutral
- For general missingness
  - minimum distance does well, NSI and SI – efface missingness
  - NSI does well in terms of the association between the clusters and the amount of missingness
  - As missingness increases NSI is more effective than SI
  - But imputation is still probably better