Multiple Imputation for Life-Course Sequence Data

Brendan Halpin
Dept of Sociology
University of Limerick*

May 2012

Abstract

As holistic analysis of life-course sequences becomes more common, using optimal matching (OM) and other approaches the problem of missing data becomes more serious. Longitudinal data is prone to missingness in ways that cross-sectional is not. Existing solutions (e.g., coding for gaps) are not satisfactory, and deletion of gappy sequences causes bias. Multiple imputation seems promising, but standard implementations are not adapted for sequence data. I propose and demonstrate a Stata implementation of a chained multiple imputation procedure that "heals" gaps from both ends, taking account of the longitudinal nature of the measured information, and also constraining the imputations to respect this longitudinality. Using the sequence data alone, without auxiliary individual-level information, stable imputations with good characteristics are generated. Using additional information about the structure of data collection (which relates to mechanisms of missingness) gives better prediction models, but imputations that differ only subtly.

Many sequence analysts proceed by cluster analysis of the matrix of pairwise OM distances between sequences. As a non-inferential procedure, this does not benefit from "Rubin's Rules" for multiple imputation in averaging across estimations. I explore ways of clustering with multiply-imputed sequences that allow us to assess the variability due to imputation. I compare the results with an existing approach that codes gaps with a special missing value that is maximally different from all other states, and show that imputation performs better.

In an example data set drawn from BHPS work-life histories, imputation of short internal gaps ($\leq$ 12 months) increases the available sample size by approximately 25 percent. Moreover, the gappy sequences have a distinctly different distribution, with higher numbers of transitions, so deletion of gappy sequences distorts the sample badly. For typical longitudinal data sets, we can expect missingness to be related to the amount of instability in the career, and to proceed without imputation will cause serious bias.

*This is an early-stage draft: please contact me for an updated version before quoting. Thanks to Rebecca Lacey whose query about MI for sequence data prompted this research. shortptex.tex,v 1.1 2012/05/02 22:14:55 brendan Exp

Contents

1 The problem and a proposed solution

This paper proposes an implementation of multiple imputation (MI) for missing values in life-course data, in the context of sequence analysis. Ever-increasing amounts of longitudinal data are available, and a growing set of tools (among which is sequence analysis) exists to handle it. However, longitudinal data is susceptible to missingness at least as much as cross-sectional. Indeed, repeated collection of data on continuous processes raises a whole set of difficulties. Life-course histories are thus prone to gaps and other forms of missing information (Halpin, 1998).

I present this work in the context of sequence analysis, which is to say the analysis of life-course trajectories as whole units, typically by analysis of pairwise distances between trajectories as measured by algorithms such as optimal matching (e.g., Abbott and Forrest, 1986; Halpin and Chan, 1998; Blair-Loy, 1999; Pollock, 2007; de Saint Pol, 2007; Martin et al., 2008; Aisenbrey and Fasang, 2010). As an approach it has advantages and disadvantages, but it is particularly vulnerable to missing data. While there are methods that deal with gaps (and I discuss one below) they have shortcomings, and the usual strategy is to discard incomplete records. Since certain sorts of trajectories are more prone to missingness (particularly those with higher instability) this is a serious distortion, as well as a potentially important loss of sample size.

There are two main issues with the use of multiple imputation in this framework. First is the longitudinality of the data: the imputation needs to be structured to take account of this, but the longitudinality also means we have very good information from which to impute. Second is that insofar as sequence analysis generally proceeds by non-inferential methods such as cluster analysis and multi-dimensional scaling, it does not benefit directly from MI's calculation of the combined variance across imputations, as per "Rubin's Rules" (Rubin, 1987). I investigate a number of ways of taking account of the variability of the imputations in a cluster analysis context.[1]

[1] The concept of discrepancy (as a partitionable measure of the variability inherent in a pairwise distance matrix) offers the potential of using inferential methods with the distances derived from

2 Standard MI inappropriate for longitudinal data

Multiple imputation is emerging as one of the best ways of dealing with data sets containing missing values. Unless the missing data is "missing completely at random" (in Rubin's terminology; abbreviated MCAR), the standard practice of analysing only complete cases will introduce bias. Similarly, discarding variables with any missing value will often throw away essential information. If the missing data is "missing at random" (MAR; that is, random given the observed variables) then imputation of the missing values using the observed data will give good estimates of the distribution of the missing data. Multiple imputation makes several draws from the predicted distribution. If for each imputation the end analysis is carried out once, we can average the results and arrive at unbiased estimates of the parameters of interest and their variability, by averaging across the analyses. This is not to create artificial data, but rather to extract the full information from the non-missing data we do have. Rubin's contribution is to point out how to do this, and that we can approximate the distribution of the unobserved data with relatively small numbers of imputations.

Where multiple variables have missing values, chained imputation is conventionally used (Royston, 2004). Typically one begins by imputing the values for the variable with the least missing using the observed data, then proceeds to the next-least-missing variable and imputes that using observed and already-imputed data, and so on. It is essential to predict using observed and already-imputed data to ensure that the case as a whole "makes sense" (though it is probably best that the predictive model is estimated on observed data only).

Longitudinal data such as lifecourse sequences present problems for multiple imputation. First, the time-series of state observations will have much higher levels of collinearity than most cross-sectional data, and if we treat the repeated observations as separate variables, very high numbers of candidate predictors. Secondly, missingness will also be serially correlated: there are "gaps" rather than single missing values. The former problem means that standard approaches to chained multiple imputation will be inefficient and unwieldy. The latter means that imputations that do not take account of the linear structure, and that gaps rather than single missing values must be imputed, will produce unrealistic imputations with spurious transitions. For instance, `AAA....BBBB` may be imputed as `AAABABABBBB` whereas a single transition such as `AAAAABBBBBB` is more realistic.

2.1 Longitudinally-appropriate recursive MI algorithm

The novelty of this paper is to propose an implementation of multiple imputation that is appropriate for life-course sequences, or more generally of discrete time series of discrete states. This involves treating the multiple variables as time-specific observations of a single state, and recognising that the gap to be imputed is composed of a related string of observations. The key requirement is thus to control the order of the chaining to respect the longitudinal structure, closing the gap from its edges, since the most important predictor by far is the immediately adjacent state. We can consider any available information

the imputed data (Studer et al., 2011).

(such as individual fixed characteristics) in the imputation model, but the best information for prediction is other observations in the same time series (and summaries thereof). It is an empirical question whether individual-level information will improve the imputation but in my experience, the information in any sufficiently long sequence is likely to make it redundant, or at least of only secondary benefit. Other information that may be valuable is that related to the time (calendar time, such as, say, employment rate, or developmental time, such as age).

The imputation technique here is motivated by sequence analysis, and I go on to consider how to exploit the multiple imputations in the context of cluster analysis of the pairwise OM distances. However, the imputed sequences are likely to be useful for other analytic approaches as well, such as hazard rate models or other models of transition rates. Hazard rate modelling may raise issues of exact timing of imputed transitions, and of what exactly constitutes the unit that is multiply imputed, but should be capable of being accommodated in the framework.

In imputing a missing value in a sequence, the best predictors are those closest to it in time. This leads us to beginning with the edge of a gap: predict the first (or last) element in the gap using the immediately adjacent value and the one at the other end of the gap (and perhaps other variables, but at least these). If the gap is five units long and we are predicting the first element, the imputation model predicts state at t using (at least) state at $t-1$ and $t+5$. Then the last (or first) element of the gap (now 1 unit shorter) is the next best estimated, predicting t by $t+1$ and $t-4$, noting that the state at $t-4$ is imputed. While the prediction model should be estimated only on observed data, the prediction must be based on already-imputed values where relevant, in order to generate a series of imputations that makes longitudinal sense.

Filling from the edges means that the longitudinal nature of the state history is respected, and that we begin the imputation with the easiest states to predict. We can implement this switching strategy by starting with the length of the longest gap to be imputed (either the longest gap in the data, or the longest gap the analyst chooses to impute, discarding cases with longer gaps) and implementing an even-first (or odd-first) rule which is followed throughout the data set. In the unlikely event that the even/odd character of gap-length is informative, and not MCAR, then it would be sufficient to assign each gap at random to even-first or odd-first. This would mean, however, estimating twice as many models: in the simple case we estimate

$$\text{for } i = 0 \rightarrow g-1 : \hat{s}_t = \begin{cases} f(s_{t-1}, s_{t-(g-i)}), & g-i \text{ if odd} \\ f(s_{t-(g-i)}, s_{t+1}), & g-i \text{ if even} \end{cases}$$

i.e., one predictive model per gap length, whereas if gaps are assigned even-first or odd-first at random, both models will have to be fitted for each gap length.

2.2 Sketching gap closure

Figure 1 outlines the procedure visually, considering a five-unit and a three-unit simulation. It assumes an even-first order, and the simplest possible imputation model, so in the first pass the last element of the five-unit gap is imputed,

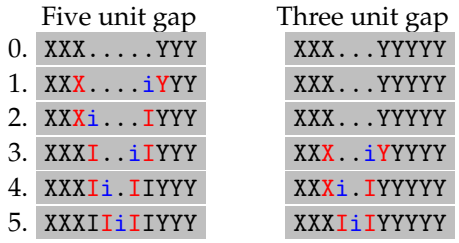

Figure 1: Sketching the process of gap closure. The left column represents a sequence with a five-unit gap, the right one with a three-unit gap (line 1). The first pass imputes the last element of the five-unit gap (even-first) but does not touch the three-unit. The second pass imputes the first of the now four-unit gap, and it is only at the third pass that the three-unit gap is touched. The procedure continues until both gaps are filled simultaneously.

using the $t-5$ and $t+1$ observations. Nothing is imputed in the three-unit gap because $t-5$ and $t+1$ are not the best available predictors for any element of this gap. In the second pass, the five-unit gap has been reduced to four units, so the first is predicted, using observed $t-1$ and already imputed $t+4$. In the third pass, both gaps now have the same length, so both are treated equivalently, except the originally-five unit gap is predicted by observations both of which have already been imputed. The procedure continues until all gaps are filled simultaneously.

2.3 Minimal model

The minimal model used above uses the immediate before and after information, in a multinomial logistic regression of the following form:

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

where one of δ_1 and δ_2 is 1, the other the (reducing) gap length. Forms other than multinomial logit may be appropriate in certain circumstances, but in sequence analysis the state space is almost always nominal. Note that the dependent and predictive variables have the same categories, so the indicator parameters for the explanatory variables have the same index as the dependent variable.

In Stata terms the looks like this, assuming the data in "long" format:

```
by id: gen last = state[_n-‘delta1’]
by id: gen next = state[_n+‘delta2’]
mlogit state i.last i.next
```

with local variables `delta1` and `delta2` assigned appropriately.

The lagging and leading states are essential to the procedure, and on their own they make for apparently good imputations. However, the model is imperfect and I consider improvements in the next section.

Note also that since these are purely predictive models, concerned not with causal relationships but with modelling joint distributions, it is as appropriate

to think about states after the gap affecting it (in a statistical sense) as states before: in this sense what happens after the gap is "history" just as much as what precedes it.

The imputed value is derived by assigning to the states at random, with probability predicted by the multinomial logistic regression. Inspection suggests that in most cases the predicted probabilities are very unevenly distributed (one close to 1.0 and the others low, or two large). That is, the predictions usually strongly reflect the prior and subsequent states.

2.4 Improving the predictive model

The baseline predictive model takes account only of immediately prior and subsequent state. This implies a uniform time-constant Markov process underlying the sequences, which is clearly not realistic. There are many ways to enhance the predictive model. Taking account of more sequence information (such as summaries of the prior or subsequent histories) is one (not Markov). Taking account of fixed individual characteristics such as gender, education or class of origin, is another (not uniform). We can allow the transition rates to change with time by including an interaction between time and the effects of the predictors, so that a gap early in the sequence will be filled differently from one late in the sequence (not time-constant). We can further take account of time by including external time-series, such as the rate of unemployment. We can model the effect of prior and subsequent state in a more detailed way by including their interaction into the model. However, immediately prior and subsequent states are the foundation of the prediction. Moreover, since the predictive model will be fitted many times in the process of the imputation, it is important that it is stable, and can be fitted across the sequences: more complex models may occasionally fail to converge.[2] In practice, taking account of simple summaries of prior and subsequent experience works very well to supplement immediately prior and subsequent state, and individual fixed characteristics make little difference, since they are already encoded in the observed part of the sequence. In section 7, I will return to this issue, to show that information about the structure of data collection can have an impact on the modelling.

3 A worked example

In this section I apply the model to a real data set, consisting of six years of labour market data for women with a birth at end of year 2. This is derived from British Household Panel Survey data, relating to the status at interview, and in the inter-wave period. The data consist of 872 full sequences, coded into four states: full-time employed, part-time employed, unemployed and non-employed, recorded monthly over six years. There are an additional 207 sequence with internal gaps up to 12 months in length. Relatively few sequences have longer gaps, but substantial numbers of cases with initial or terminal missings are ignored. There is no overriding reason to ignore in ital or terminal gaps, though one may wish to limit them to shorter spans as they have

[2]The Stata implementation has a fallback procedure, such that if a complex model fails to converge, a less complex one can automatically be estimated instead.

only half the relevant information. I exclude them here primarily because they need to be modelled slightly differently, as they miss either prior or subsequent information.

3.1 Predictive models

The predictive model chosen for the imputation needs to be substantively meaningful, and stable. In work not reported in detail here, I have examined the performance of a range of models for a variety of gap lengths, in terms of their overall fit (particularly in terms of the likelihood ratio test on adding terms to the model, but also in terms of the pattern of agreement of predicted probabilities).

The simplest model used just takes immediately prior and subsequent states into account: `mlogit state i.last i.next`. It seems attractive to include the interaction between these states: `mlogit state i.last##i.next`. However, some combinations are rare, and in general this model was very unstable.

It was also evident that average transition rates changed strongly with time: in the nature of this data set, transition rates peak around the period of the birth. This can be modelled by interacting quadratic time with the prior and subsequent states.

Given that the sequences do not arise from Markov processes it is important to include information on history and future. This is done by calculating the proportion of time in each of the categories before and after the gap.

While there is a strong conceptual case for including individual fixed characteristics (e.g., class of origin, age at t_0), in practice these seemed to have relatively little effect. This is probably because their effect on the states in gap is the same as their effect on the observed states, and this is already incorporated. It was considered going further in this direction by fitting a random-effects multinomial logit, to account for all individual heterogeneity; this is however too computationally demanding, and would require computing time measured in days, as well as being prone to convergence problems.

The working model took the following form, as a Stata expression:

```
mlogit state i.next##c.t##c.t i.last##c.t##c.t ///
             before1 before2 before3 ///
             after1 after2 after3, iterate(40)
```

Note the `iterate(40)` option: not all models converged within 40 iterations, so the code included a test for convergence with an automated fallback to a simpler model, in order that the chained imputation sequence completed. In terms of the sorts of imputations made, there was relatively little difference apparent when comparing this model to simpler ones, at least to superficial inspection of the imputed sequences.

Figure 2 shows some example imputations, for six representative sequences. Each block represents an observed sequence with ten imputations; the grey in the top row is the gap (some sequences have more than one gap: this is typical). It is clear from inspection that the imputation behaves in general as expected. Gaps bracketed by a single state are usually filled in with that state (none shown in the example, but reasonably common in the complete data). Gaps bracketed by two states usually have a single transition between the two, with the transition time being random. Some extra transitions or third states are imputed, but these are less common, if more likely with longer gaps.

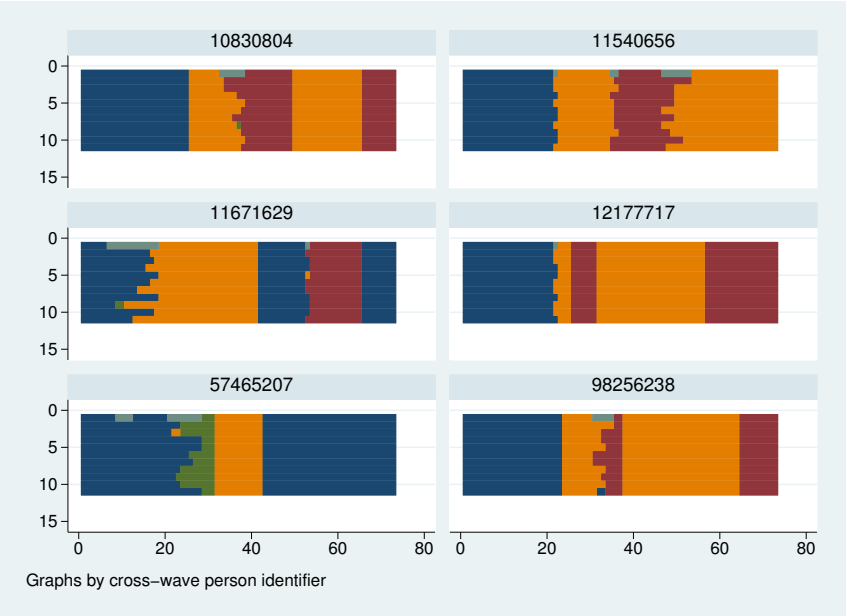

Figure 2: Examples of imputations: gappy sequences with 10 imputed copies. The grey sections in the top row of each block represent gaps; in the rows below we can see ten examples of what the model imputes.

At present, only visual inspection of the imputations has been carried out, but it would be desirable to analyse them more systematically, and to relate this to a formal comparison of the predictive models.

To summarise the experience with imputations within this data set, the model used is reasonably complex but does not take fixed individual characteristics into account. A certain amount of models fail to converge, so the fall-back strategy is important but is used relatively rarely. The imputations produced have (at least superficially) good or verisimilar characteristics, in terms of randomisation of transition points, and occasional introduction of third states. Quite importantly, imputation increases the data set from 872 to 1079 cases, an increase of almost 24 percent. As we will see below, these additional cases are disproportionately "interesting" sequences with multiple spells.

We now move on to the issue of how to benefit from these multiple imputations in analysis.

4 How can we exploit MI of lifecoure sequences?

Multiple Imputation normally proceeds by stochastic modelling of the multiply imputed data sets. Parameters are simply averaged across the replicated data sets, yielding estimates of the true value that are no longer biased by missing data, subject to assumptions about the nature of the missingness. The variance of the estimate is also calculated across the imputations, in a slightly more complicated way (see for instance Royston, 2004, p. 273). However, in the usual workflow of sequence analysis, the next stage is the generation of pair-

wise distances by OM (or other measure), followed by cluster analysis. These are methods which are not usually framed in inferential terms, so there is no ready analogue of standard error. So how do we benefit from the information carried through the multiple imputations?

We can consider a number of strategies. The simplest would be to take the average distance between the imputed sequences and all other sequences, and treat that as the definitive distance set for that sequence. That is neat and efficient in that it yields a single measure, but it completely obscures the variability inherent in MI.[3] This method is considered further, below (section 6.1). Another simple approach would be to cluster the imputed data sets separately, and compare the pattern of cluster membership, particularly for different imputations of the same sequence. However, clustering is relatively unstable, in the sense that small changes in the data will sometimes lead to disproportionate changes in the cluster solution, so it can be hard to compare across imputations – in particular, it is hard to deal with assessing the affect of different imputations on a sequence's membership of a set of clusters if the set of clusters changes as well. This instability makes it preferable to pool the imputed data sets and to conduct the cluster analysis on the combined sequences. This yields a single solution, and we can then investigate to what extent different imputations of the same sequence end up in different clusters. The downside of this approach is that it becomes memory intensive and slow, since the memory and time requirements are in theory approximately $O(N^2)$ (i.e., scale with the square of the number of cases). Moreover, Stata has a hard limit of 11,000 rows for matrices.

5 Analysis

We therefore proceed by conducting cluster analysis on the pairwise distances from the pooled data. There is no particular issue in generating the pairwise distances using OM. Similarly, clustering using Ward's linkage raises no conceptual issues with the pooled data. In particular, if there were no imputed cases, but simply a fixed number of repetitions of each observed sequence in the pooled data, we can expect the same cluster solution as if each sequence were represented once only.[4]Thus, the extent to which imputed sequences' cluster membership varies is a measure of the uncertainty in the imputation.

5.1 Pairwise distances

While there are no conceptual issues involved in calculating optimal matching distances with imputed data, there is a practical one: calculating pairwise distances is also $O(N^2)$. However, there is no reason to calculate duplicate distances, so the extra burden only applies to imputed sequences. My `SADI` add-ons for Stata[5], `SQOM` for Stata, and `TraMineR` for R all deal with duplicate

[3]This approach could be linked to the discrepancy measure (Studer et al., 2011), which is an analogue of variance, applied to pairwise distance matrices. Using this approach, the variability becomes meaningful again. It is intended to explore this in future research.

[4]This can be verified empirically.

[5]See http://teaching.sociology.ul.ie/sadi/, or do `net from http://teaching.sociology.ul.ie/sadi/` followed by `net install sadi` within Stata.

sequences automatically (`SQOM` is, however, too slow for large data sets).

There are distance measures available other than OM. OM is the default in the sequence analysis literature. It is oriented to strings of tokens and is good at recognising similarity at different locations. However, in this context some characteristics of OM will tend to smooth differences between replications (for instance, it will be relatively insensitive to small changes in the transition point). Other measures might have different consequences. Hamming distance would possibly increase the perceived difference between imputations, and spell-sensitive measures such as time-warp edit distance TWED (Marteau, 2008) and Elzinga's X/t approach (Elzinga, 2005, 2006) will be particularly prone to treating imputations with a third state as distinctive.[6]

5.2 Clustering

Clustering with Ward's method is the typical sequel to calculating distance measures. Its primary purpose is to generate an empirical typology of the sequences. Ward's linkage performs well: in my experience it is more likely than other readily available linkages to yield a solution with reasonably even distribution of cases across the clusters; many of the other linkages tend towards one huge cluster and many tiny ones.

Clustering a large data set is slow, and as noted above, Stata imposes a strict limit of 11,000 cases. Thus we are limited in the number of replications we can use. However, for a data set of up to 1,100 cases, we can have 10 replications. Considering, however, that in this pooled data set that there are very high numbers of duplicates, it would be desirable that the clustering could work with weighted data. For the fully observed sequences, they could enter with a weight equal to the number of replications, and only the imputed sequences would increase with the number of replications. Were this possible, much larger numbers of replications would be feasible (and clustering with any duplicates would be significantly faster). Ackerman et al. (2011) demonstrate that clustering weighted data is possible for many linkages (including Ward's), but I have not been able to find an implementation.

However, for the present data set, with a size of 1,079 including imputed sequences, Stata will accommodate up to ten replications. On an elderly PC (3MHz, non-SMP Stata) this data set requires approximately four hours to run the cluster analysis. A five-replication version completes in less than 30 minutes, suggesting that the process is worse than $O(N^3)$. This suggests that it would be worth investigating alternatives such as R's `fastcluster` library, which claims to be $O(N^2)$, and is not subject to the 11,000 case limit. However, for present purposes Stata does the job, and Figure 3 shows the eight-cluster solution that results.

5.3 Variation in cluster membership

The cluster solution in Figure 3 would normally be the product of interest, but here our focus is not so much on the classification of trajectories as on the vari-

[6] A release of `SADI` containing a larger range of distance measures is in preparation. As well as OM, it will provide Hamming distance, Lesnard's dynamic Hamming, Hollister's localised OM, Halpin's duration-adjusted OM, TWED and a version of Elzinga's duration-weighted combinatorial measure.

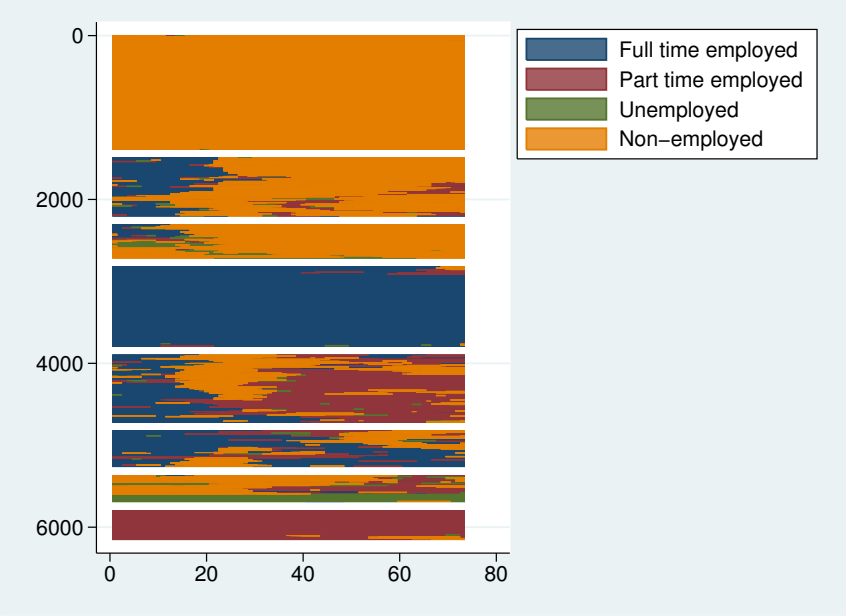

Figure 3: The eight-cluster solution, for five replications

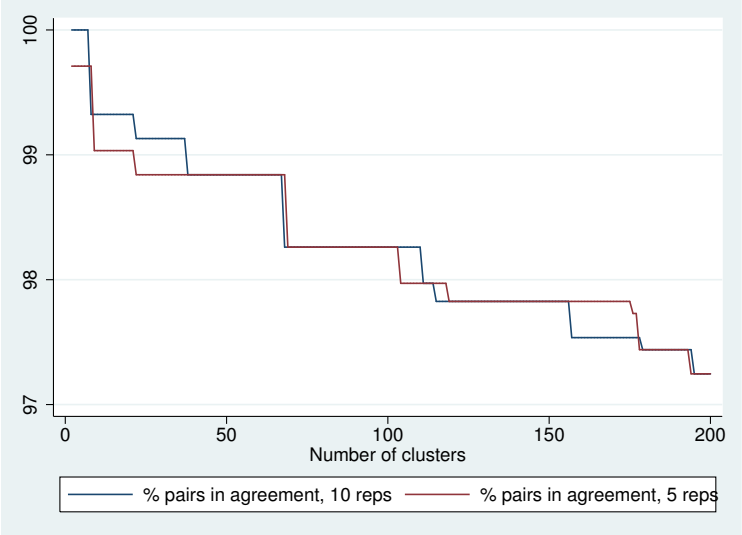

Figure 4: Agreement between imputed sequences

ability of the imputed sequences. That is, for a given cluster solution, and a given gappy sequence, to what extent are different imputed versions of the sequence assigned to different clusters? A simple way to measure this agreement is to calculate for each base sequence, how many pairs of imputed sequences disagree. Given r replications of a sequence, there are $\frac{r(r-1)}{2}$ pairs, and we can create an index of the proportion of pairs where both sequences fall in the same cluster. Figure 4 summarises this quantity across all imputed sequences, for five and ten replications, for cluster solutions between 2 and 200 categories. Clearly, the more clusters in the solution, the easier it is for two imputations of the same sequence to fall in different clusters. From this figure we see a number of features. First is that as the number of clusters rises, so does the disagreement, but at a relatively slow rate. Second is that the picture is quite similar for five compared with ten replications. Third, as the number of clusters increases, many of the splits have no effect on the index. But perhaps most important is that fact that even at very high numbers of clusters, more than 97 percent of imputed sequences are in agreement. Inspection of the data shows that some of this high level of agreement is due to exact agreement between imputations: naturally, given the discrete state space and the fact that some gaps are short, some imputations are exactly identical. What is probably more important, however, is that many of the imputations differ in small ways that OM will discount, such as having a small difference in the timing of the imputed transition. Coupled with the fact that the imputed proportion of the sequence is relatively small (sequences are six years long, and the longest gap is one year, though sequences may have more than one gap), this means that variation among imputations is often very small. This is an important conclusion, but one that depends on the specific data set: imputation adds relatively little uncertainty to the analysis. For other data sets, with more missing data, or a bigger state space, this conclusion might not hold, but looking at agreement measured in this way will be informative.

5.4 Comparing imputed with fully-observed sequences

Given the relative stability of the imputed sequences, the next question that arises is how similar the incomplete sequences are to the complete ones. Figure 5 shows the distribution of complete and incomplete sequences across the eight-cluster solution (see also Table 1). It is immediately clear that simple clusters of simple sequences (particularly clusters 1, 4 and 8, composed mostly of trajectories entirely in a single state) are very predominantly composed of completely observed sequences. In contrast, clusters composed of sequences that experience multiple transitions have much higher proportions of imputed sequences (in particular clusters 2, 5 and 6). Given what we know about the relative stability of the imputations, we can say that this is a true feature of the incomplete sequences and not an artifact of the imputation. That is, incomplete sequences are much more likely to display high rates of transition in their observed portions, than are complete sequences. There are probably many reasons for this, but a major one is that individuals with unstable or complex life situations are more likely to be hard to contact for interview, and will have more opportunity for measurement error to enter when they are interviewed because they have more information to report (dates of transitions,

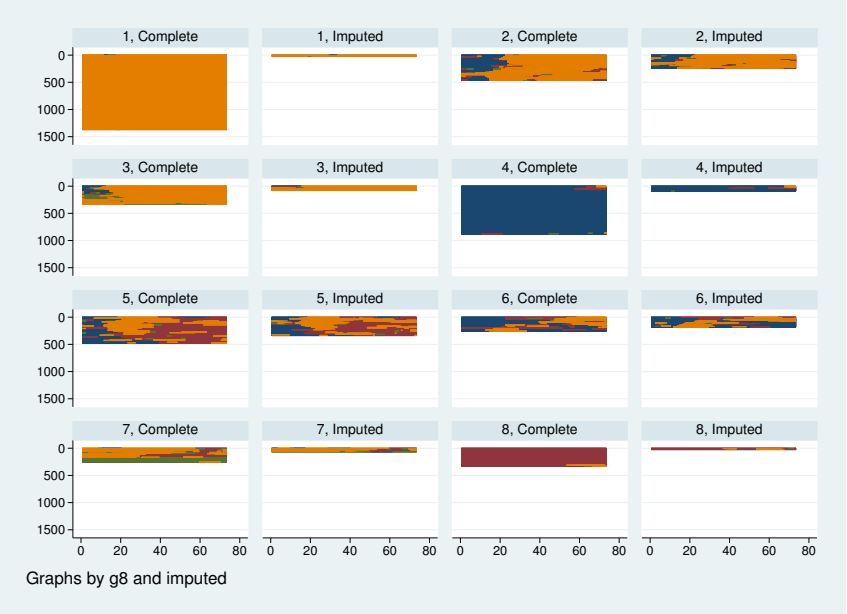

Figure 5: Comparing complete with imputed sequences, eight-cluster solution, five replications

etc). It may also be the case that there is measurement error in some of the completely-observed simple sequences, i.e., that gaps in the record that represent an unobserved spell in another state are "papered over" by reporting a too-early date for the return to the main state.

5.5 Reviewing the imputation exercise

To recap the imputation exercise, the predictive models seems to be generating plausible and stable imputations, which in turn are clustered very similarly, with in excess of 97 percent in the same clusters. Very importantly, in this data set imputing internal gaps up to 12 months long increases the available sample size by a quarter. Moreover, incompletely observed sequences are distributed significantly differently from completely observed sequences, being much more likely to be complex, multi-transition sequences. In this sort of data, quite aside from representativeness issues, there is a problem that the space of these "interesting" sequences is fairly sparse, so an increase in their presence is particularly welcome, and facilitates the clustering of complex sequences. Representativeness is important too, of course, so the inclusion of these sequences reduces bias substantially. At least impressionistically, therefore, multiple imputation of lifecourse sequences seems to be an attractive option, and is much better than the basic alternative of modelling only complete sequences.

Table 1: Distribution across 8-cluster solution of complete and imputed sequences, five replications, percentages in parentheses

Cluster	Complete	Incomplete	Total
1	1,370	15	1,385
	(31.42)	(1.45)	(25.67)
2	465	242	707
	(10.67)	(23.38)	(13.10)
3	325	80	405
	(7.45)	(7.73)	(7.51)
4	880	90	970
	(20.18)	(8.70)	(17.98)
5	480	338	818
	(11.01)	(32.66)	(15.16)
6	255	180	435
	(5.85)	(17.39)	(8.06)
7	255	65	320
	(5.85)	(6.28)	(5.93)
8	330	25	355
	(7.57)	(2.42)	(6.58)
Total	4,360	1,035	5,395
	(100.00)	(100.00)	(100.00)

6 Alternative approaches

Cluster analysis of the pooled data is one approach to using the multiple imputation data. Other approaches are possible. I consider two here: using the average distance between the multiple imputations and other sequences, and using OM with gaps marked as missing values.

6.1 Averaging distances

Averaging distances is a way of reducing the multiple imputations to a single entity. It is not possible to average the imputed sequences to a single sequence, but we can find the average distance between the set and each other sequence. Thus we end up with a single set of distances for each incomplete sequence. This has the advantage of not inflating the data set, but loses all sight of the variability of the imputations. In principle we should end up with a very similar picture to the pooled cluster analysis, except that the imputed sequences now belong to exactly one cluster. The clustering will not be exactly the same, however, since the points implied by the averaged distances are not the same as those implied by the multiply imputed distances. Small changes in the data will tend to lead to moderate inconsistencies in the cluster solution. We can see this by comparing the cluster membership of the sequences under this analysis, with that under the pooled analysis (for five replications). Table 1 cross-tabulates the MI solution with the average-distance solution (with the latter permuted to give the largest possible agreement; as clusters have no labels, clusters in two different classifications can only be identified on the basis

Table 2: Comparing the average-distance and the pooled solutions, five replications

MI solution	Average-distance solution								
	1	2	3	4	5	6	7	8	Total
1	1,385	—	—	—	—	—	—	—	1,385
2	5	655	—	—	2	10	35	—	707
3	90	—	315	—	—	—	—	—	405
4	—	—	—	970	—	—	—	—	970
5	—	30	—	—	678	110	—	—	818
6	—	15	—	100	—	320	—	—	435
7	—	20	—	—	—	—	300	—	320
8	—	—	—	—	—	—	—	355	355
Total	1,480	720	315	1,070	680	440	335	355	5,395

of shared membership). For one run, we get agreement measured by the Adjusted Rand Index of 0.866 (Vinh et al., 2009), and a κ_{max} index of 0.908 (Reilly et al., 2005).[7] These are high levels of agreement, but are far from perfect. As we can see from the table, almost 8 percent of cases are off the diagonal. It should also be noted that these figures vary if we impute another set of cases (in other words, five imputations is likely too low for stability).

As an index of the instability of clustering, approximately 4 percent of completely observed sequences are placed in different clusters, as a result of the presence of the averaged or multiply imputed sequences. Over 20 percent of the imputed sequences are assigned to different clusters. From all this I conclude that while averaging is convenient and may be a way of accommodating imputation in larger data sets, it is not equivalent to clustering the multiply imputed distances.

6.2 A second alternative: missing as special category

As I stated at the beginning, there are other methods for dealing with gaps in sequences, at the level of defining the pairwise distances. These completely avoid the issue of imputation, and it is worth exploring them. One strategy, which is proposed in the TraMineR documentation (Gabadinho et al., 2009, section 9.4.5), is to treat missing states as belonging to a special category, and to assign substitution costs to this category. An undesirable consequence of this suggestion, which is explicit in the documentation, is that missing states are regarded as identical, that is, a missing state in one sequence counts as a perfect match with a missing state in another. I adapt the approach to avoid this, and justify it as follows: Missing means the state is unknown but belongs to the state space. We can therefore assign the maximum substitution cost to the substitution between missing and any known state, as a ceiling. Similarly, we can assign the maximum substitution cost to the match between missing and missing, on the grounds that there is no more reason to assume that the two

[7] The ARI compares two unlabelled classifications and indexes the extent to which any pair of cases is in the same (or different) groups in one classification is in the same (or different) groups in the other; 1.0 implies perfect agreement. The κ_{max} is Cohen's Kappa for the permutation of the classification that maximises it.

Table 3: Missing as a special state, versus MI: comparison of cluster solution

MI solution	"Missing as state" solution								
	1	2	3	4	5	6	7	8	Total
1	1,385	—	—	—	—	—	—	—	1,385
2	—	440	215	—	12	35	5	—	707
3	15	—	390	—	—	—	—	—	405
4	—	—	—	960	—	10	—	—	970
5	—	—	—	65	603	145	—	5	818
6	—	—	—	70	5	345	—	15	435
7	—	5	—	—	—	—	315	—	320
8	—	—	—	—	—	—	—	355	355
Total	1,400	445	605	1,095	620	535	320	375	5,395

unobserved states are a match than in the case with one unobserved and one observed state. This requires a substitution matrix with a non-zero value on the diagonal, which is counter intuitive but can be accepted as long as we view the missing category as "unobserved" rather than as an identified extra category. Depending on how the OM algorithm is programmed, this special substitution matrix will propagate correctly into the calculations (it works with `SADI` but I have not tested `TraMineR`). The net result of the strategy is that a sequence with missing values has distances calculated that are greater than or equal to the distances that would be calculated if the sequence was fully observed.

Clearly this has a different rationale to multiple imputation, maximising the difference due to missing data rather than replacing it with plausible values. However, it offers a useful benchmark. The question is, how often does this approach put sequences in different clusters from the MI approach? As we see from Table 3 there is a relatively high level of disagreement (11 percent of cases off-diagonal, ARI 0.832, κ_{max} 0.868). Among the incomplete cases, 22 percent disagree.

From this we see that while the approaches have the same effect in including sequences that would otherwise be discarded, the results are not equivalent. Intuitively, given that the imputed sequences are usually quite plausible, in the sense of being based on observed data, whereas the missing-as-state distances are maxima, intentionally biased upwards, it is likely that the MI distances are in general more accurate than the missing ones.

7 A caveat: is there more to missingness?

To return to the issue of the prediction model, I want to introduce a final qualification. The prediction model needs to respect the mechanisms by which observations become missing, and with longitudinal data we can sometimes say more about this process. In particular, if missingness applies to spells of time, rather than to states at particular dates, it is likely that we can improve the model. A typical way in which a gap can develop in panel data is due to the interviewee missing an interview (at year Y). If the account of the life-history given at the interview at year $Y + 1$ does not extend back to the date

of interview at Y, there will be a gap in the record, starting immediately after interview Y. In this case the structure implies that while there was almost definitely a transition at the end of the gap, there is no particular reason for supposing the gap begins with a transition.

The details depend on precisely how the data were collected, and what information about this persists through to the final data set, but in general if information about the structure of data collection is available it may improve the prediction model in significant ways. In the case of the BHPS, the key information that propagates into the final data set is whether interview took place in that month, and whether the month is explicitly accounted as the first or last month of a spell. Adding this information to the model described in section 3.1 makes a big difference as measured by the likelihood ratio test. However, the differences it makes to the imputations are subtle, to say the least. Visual inspection suggests that there may be more transitions at the edges of gaps, but the distribution of states and the numbers of transitions imputed doesn't seem to change materially. This improvement in imputation is probably much more relevant to methods that are sensitive to duration and to transition rates, such as hazard rate modelling.

More work is needed on this issue, in particular to make a more formal comparison of the imputations, and to examine them in hazard modelling frameworks.

8 Conclusion

Resolving the problem of gaps in sequence data is essential. Longitudinal data is particularly subject to missingness, so discarding incomplete cases can be particularly costly in terms of reduced sample size and loss of representativity. And as we have seen, not only are incomplete sequences distributed differently from complete, but they are concentrated among the "interesting", high-entropy sequence where there is a lot going on. From a cluster analysis point of view these sequences occupy a part of the trajectory state space that is hard to cluster because it contains so much variability, and therefore sparseness is a particular problem.

As a response to missingness, multiple imputation of lifecourse sequences seems to work very well. In the example used, the amount of information in the sequences is enough to impute the gaps with relatively little variability, though for data with longer gaps or a larger state space, this may not be so true. The formal structure of the imputation, where gaps are "closed" incrementally from the edges, with prior and subsequent state having the key predictive role, seems to be effective. The distribution of the imputations is as one would expect, on the basis of visual inspection, though more formal testing is planned. While prior and subsequent state work very well on their own, it is evident that more is needed in the predictive model: the implied constant uniform Markov process is clearly not realistic. Improving the model by taking account of time, and prior and future "history", makes a big difference to the fit of the model, but not a dramatic difference to the imputations (again, a more formal examination of the imputed values is intended). Whether it is valuable to include fixed individual characteristics is a question that has not been resolved, though it seems that the additional information they would bring in,

having taken account of the observed part of the sequence, is likely to be small.

Clustering as a sequel to multiple imputation is unusual, and loses some of the benefit of averaging inference across imputations. Further, it is memory and computationally intensive, which is an issue that limits the number of replications that can be dealt with. However, as a practical issue cluster analyses of reasonable numbers of replications can be dealt with in acceptable time, using Stata's default capabilities. The hard limit of 11,000 cases imposed by Stata's matrix system is a serious issue, and it seems necessary to explore other software, particularly for larger data sets or higher numbers of replications. Nonetheless, the strategy of pooling imputations, and clustering the combined data set, is effective, particularly when agreement between pairs of imputations can be tracked.

While clustering is an obvious thing to do with pairwise distance data, it has its limits and its instabilities, and it is worth looking to other techniques: discrepancy analysis (Studer et al., 2011) has particular promise in this respect, and is an avenue for further research. Looking at the average distance across imputations connects with this perspective.

Two alternative approaches were examined: reducing the distances of the multiple imputations to an average and thus reducing the imputed data set to size N, and treating missing as a special maximally different, non-self-identical state. The main advantage of the former is that we get a multiply imputed measure of the location of the incomplete sequences in the sequence space, and can proceed without the costs associated with large sample size: for data sets too large to cluster in their replicated format this offers some potential. However, since it loses sight of the variability of the imputations it is less attractive than clustering the replicated data, where that is possible. The second alternative, that of treating missing as a special state, also has some value, in providing a benchmark in that it includes the missing data in the analysis, but without any imputation. However, it includes the incomplete sequences in a somehow grudging way, and their distances to all other sequences are systematically biased upwards. Insofar as the imputation model is sound, the multiply imputed distances (averaged or as a distribution) are less biased.

At any rate, the standard practice of dropping incomplete cases is in sequence analysis, as in other forms of data analysis, clearly inferior.

References

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, XVI(3):471–494.

Ackerman, M., Ben-David, S., Brânzei, S., and Loker, D. (2011). Weighted clustering. *ArXiv e-prints*, abs/1109.1844.

Aisenbrey, S. and Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research*, 38(3):420–462.

Blair-Loy, M. (1999). Career patterns of executive women in finance: An optimal matching analysis. *American Journal of Sociology*, 104(5):1346–1397.

de Saint Pol, T. (2007). Le dîner des français : un synchronisme alimentaire qui se maintient. *Économie et Statistique*, 400:45–69.

Elzinga, C. H. (2005). Combinatorial representations of token sequences. *Journal of Classification*, 22(1):87–118.

Elzinga, C. H. (2006). Sequence analysis: Metric representations of categorical time series. Technical report, Free University of Amsterdam.

Gabadinho, A., Ritschard, G., Studer, M., and Müller, N. (2009). Mining sequence data in R with the TraMineR package: A user's guide for version 1.2. Technical report, University of Geneva.

Halpin, B. (1998). Unified BHPS work-life histories: Combining multiple sources into a user-friendly format. *Bulletin de Méthodologie Sociologique*, (60).

Halpin, B. and Chan, T. W. (1998). Class careers as sequences: An optimal matching analysis of work-life histories. *European Sociological Review*, 14(2).

Marteau, P.-F. (2008). Time Warp Edit Distance. *ArXiv e-prints*, 802.

Martin, P., Schoon, I., and Ross, A. (2008). Beyond transitions: Applying optimal matching analysis to life course research. *International Journal of Social Research Methodology*, 11(3):179–199.

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A*, 170(1):167–183.

Reilly, C., Wang, C., and Rutherford, M. (2005). A rapid method for the comparison of cluster analyses. *Statistica Sinica*, 15(1):19–33.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241.

Rubin, D. (1987). *Multiple imputation for non-response in surveys*. John Wiley and Sons, New York.

Studer, M., Ritschard, G., Gabadinho, A., and Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40(3):471–510.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.