Alignment, clocking, and macro patterns of episodes in the life course

Tim Riffe*[1] and Andrés Castro[1]

[1]Max-Planck-Institute for Demographic Research

September 29, 2019

Abstract

Background Individuals are often either observed or modeled as passing through a sequence of discrete states. These are usually either simplified into transition probabilities for Markov-derived aggregate statistics, or else retained for pattern and group detection using sequence analysis. Markov-derived statistics are of limited scope (moment stats), and sequence analysis doesn't typically lead to heuristic understanding of macro patterns.

Objective We broaden the scope of aggregate patterns and summary indices that may be calculated from trajectory data, including trajectories generated from Markov models. For example, one might calculate the time-since-event or time-to-event pattern of episode duration.

Methods We introduce the concepts of clocking and alignment as a new framework for generating novel statistics from trajectories.

Data We use different data to demonstrate concepts and give example applications. We use published transition probabilities (originally derived from US HRS data) to simulate discrete trajectories of employment states. We will use fertility and union trajectories derived from Colombian DHS data for example applications. We will also have health applications from either directly observed or simulated trajectories, tbd.

Results We demonstrate several new measures in the areas of health, family, and labor demography.

Conclusions We generate several new patterns and measures in the areas of health, family, and labor demography. An R package is presented to facilitate experimentation with these operations.

1 Introduction

There is a void between the methodological approaches of Markov statistics and sequence analysis. Markov-generated quantities, like trends and levels, are useful metrics but such information does not necessarily lead to an understanding of processes or of typical experiences. The age-structured data that underlie such Markov calculations are surely appreciated for their articulated and often-regular patterns, but (i) the estimation of such rates already blurs over the features of underlying life trajectories, and (ii) such age patterns serve the objective statistic. Insofar as sequence analysis retains and reasonably typifies life trajectories it might be used to infer processes and identify new patterns. We propose a two-part framework to extract patterns hidden within trajectory data. Such patterns might be age-like patterns of novel prevalence, state-episode-occupancy time measures, or they may be used to derive new rates. Using this framework, we aim to zoom in on demographic patterns that emerge at various stages of the life course, and so describe a given demographic phenomenon (state) from a variety of perspectives.

We first define *clock* measures, a way to inscribe time, order, prevalence, or other measures into individual life trajectories. This step is analogous to defining a rewards matrix in multistate Markov models (see e.g. Caswell and Zarulli 2018), and the ends are not entirely dissimilar. As an example of a concrete Markov link, Dudel and Myrskylä (2017a) defines a matrix algebra approach to estimate the expected number of episodes of a given state that individuals experience in a given multistate world. Taken together with the total expected state occupancy time, one can infer an average episode duration.

*riffe@demogr.mpg.de

And there are both clunky and elegant approaches close at hand to derive an age pattern of expected episode duration. Such measures (even that hasn't been done before) would already add insight to demographic processes. Clock measures are much more flexible than this, and enable the researcher to decompose expected episode durations into expected time spent and left within episodes. Further one can visualize full distributions of these and other prevalence or episode order statistics.

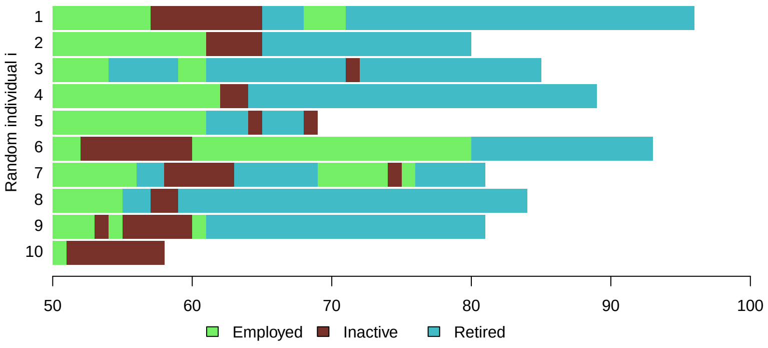
Second we define *alignment* operations, which shift trajectories to have synchronous timing with respect to a specified state-episode. Researchers already do similar things: for example Iacobelli and Carstensen (2013) propose a flexible use of time-since-event scales, and Riffe et al. (2017) define flexible Lexis spaces in which life lines are aligned both on birth and on death, and this has been used to reveal hidden health patterns (Riffe et al. 2016) and pathways (Potente and Monden 2018, Raab et al. 2018). We here propose more flexible alignment procedures, which allow trajectory synchronization on the start or end of a specified episode (e.g., first, last, longest).

In combination, clock and alignment operations open a large space for the derivation of demographic macro patterns. In the following sections we give concrete examples to illustrate these two steps. We end with a few suggestive macro patterns. At present, or examples pertain to labor demography, but in a later stage, this manuscript will include examples from different domains of demography including family, fertility, and health. We here use simulated life trajectories, although in the final manuscript we will use both simulated and observed trajectories. Work shown here is fully reproducible, and we also offer an `R` package, `Spells`, which enables flexible clock and alignment operations, and will in the future play well with other popular time series and sequence analysis packages, such as `TraMineR`.

2 Data

To demonstrate concepts, we simulate trajectories from a published transition matrix (Dudel and Myrskylä 2017b). This matrix refers to black females aged 50-100 in 1994, and it contains age-structured transition probabilities for movements between employment, inactivity, and retirement, as well as mortality from these three states. Simulation is done using the `rmarkovchain()` function from the `R` package `markovchain` (Spedicato 2017). A glimpse of the first 10 randomly generated individuals is shown in Figure 1. These ten individuals will be recycled in all of the following data manipulations used to demonstrate concepts. All aggregate calculations of age patterns (and so on) are based on a simulated population of 10000 trajectories starting in employment at age 50.[1]

Figure 1: Ten randomly generated state sequences from the 1994 transition matrix of black females (Dudel and Myrskylä 2017b)



3 Clocks

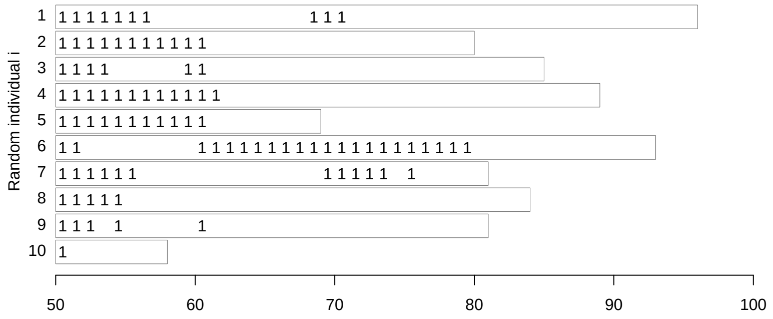
3.1 A binary trajectory matrix gives prevalence

Standard calculations of prevalence typically proceed by imputing reference states with 1s (with 0s elsewhere) and taking column means over survivors in each age. Figure 2 shows such a data construct, where

[1] We appreciate the irony that the trajectories used here came from an age-stage Markov model, which means that the diversity of patterns we derive ought to give even more food for thought.

the state sequence matrix has been converted to a binary matrix, with 1s for employment episodes, 0s for other living states (shown blank). Typically one might impute NA values in dead states for this sort of calculation. Operations on objects such as this can yield age patterns of prevalence or expectancies, for example. This is not what we call a clock, but this data construct illustrates the setup. As the number of simulated trajectories increases, the resulting age pattern of prevalence will approach the values in the respective column of the so-called fundamental matrix in a Markov approach.

Figure 2: Binary imputation of employment spells



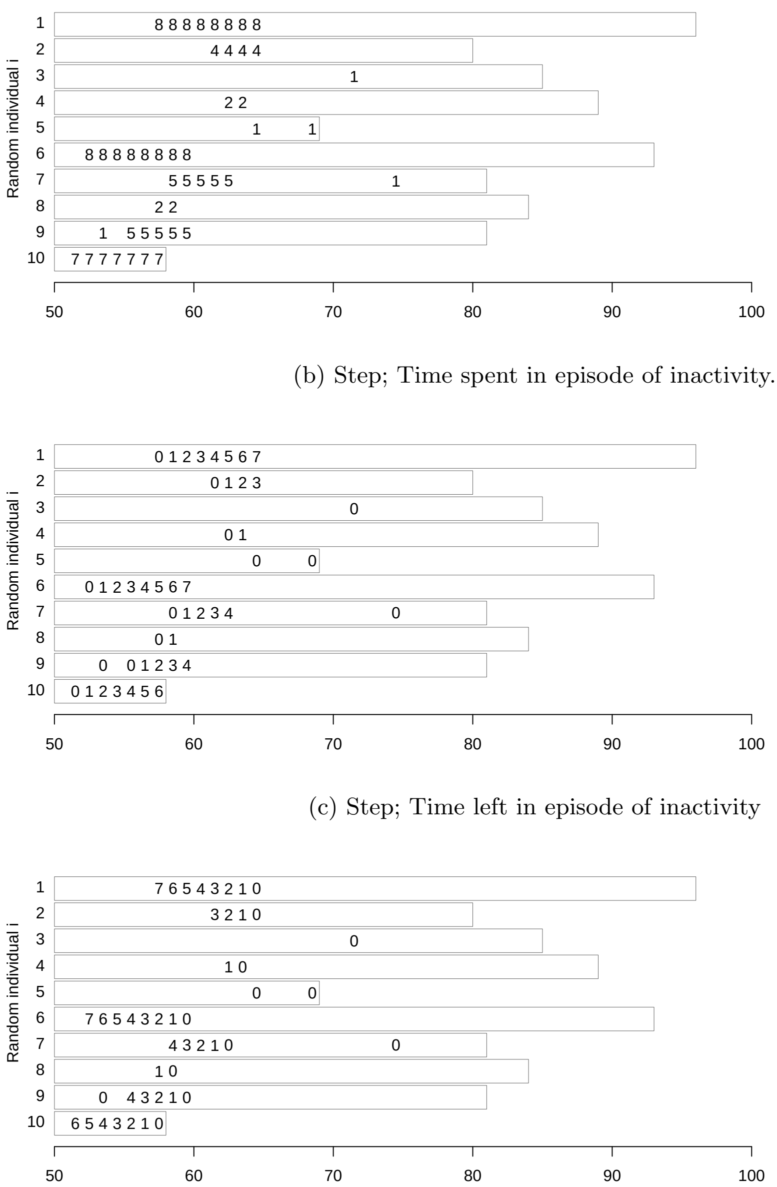
3.2 Duration, step, and order clocks

To derive measures other than prevalence, we simply change the 1s to other values. For example, if to calculate an age-pattern of spell duration, instead impute time steps with episodes with values equal to the total episode length (Fig. 3a). Column means of the resulting object would give the average episode duration conditional on being in any point of an episode. If instead one wanted to condition on episodes starting (ending) in each age then impute the same values in only the first (last) time step within each episode (not shown). One may also wish to calculate time spent or left in the state episode, per Fig. 3b or 3c [2]. Episodes can also be imputed with other markers, such as episode order, as in Fig. 4 for the case of employment spells, or episode fractions.

There is room for creativity in defining clock measures such as these, and we encourage experimentation along these lines. Clock measures are then aggregated in some way. In these examples, *value* alignment is with respect to episodes, but *aggregation* alignment is still structured by age, such that statistics across individuals in an array produce age patterns. However, one may wish to synchronize trajectories in ways other than time since birth.

[2] In practice we increment values by $\frac{1}{2}$ for mid-state clocking.

Figure 3: Inactivity spells from Figure 1 are imputed with different duration count variables. It's probably better to add $\frac{1}{2}$ to the displayed *running* values.

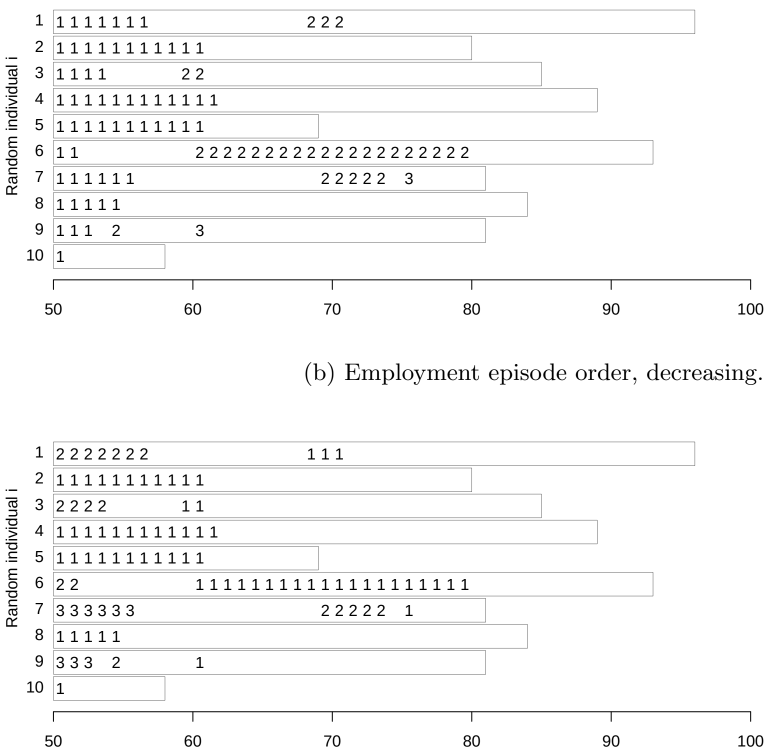


(a) Static; Total episode duration of inactivity.

(b) Step; Time spent in episode of inactivity.

(c) Step; Time left in episode of inactivity

Figure 4: Employment episodes from Figure 1 are imputed with order count variables.



(a) Employment episode order, increasing.

(b) Employment episode order, decreasing.

3.3 Alignment

Episodic *clock* values are aggregated according to some structuring criteria. In all previous figures, the structuring criteria was chronological age, which is how data were generated in the first instance. To introduce a term, the sequences in these figures are *left-aligned* on the event of birth. This is the most common default alignment in social and medical sciences, but other choices may be more compelling for particular questions.

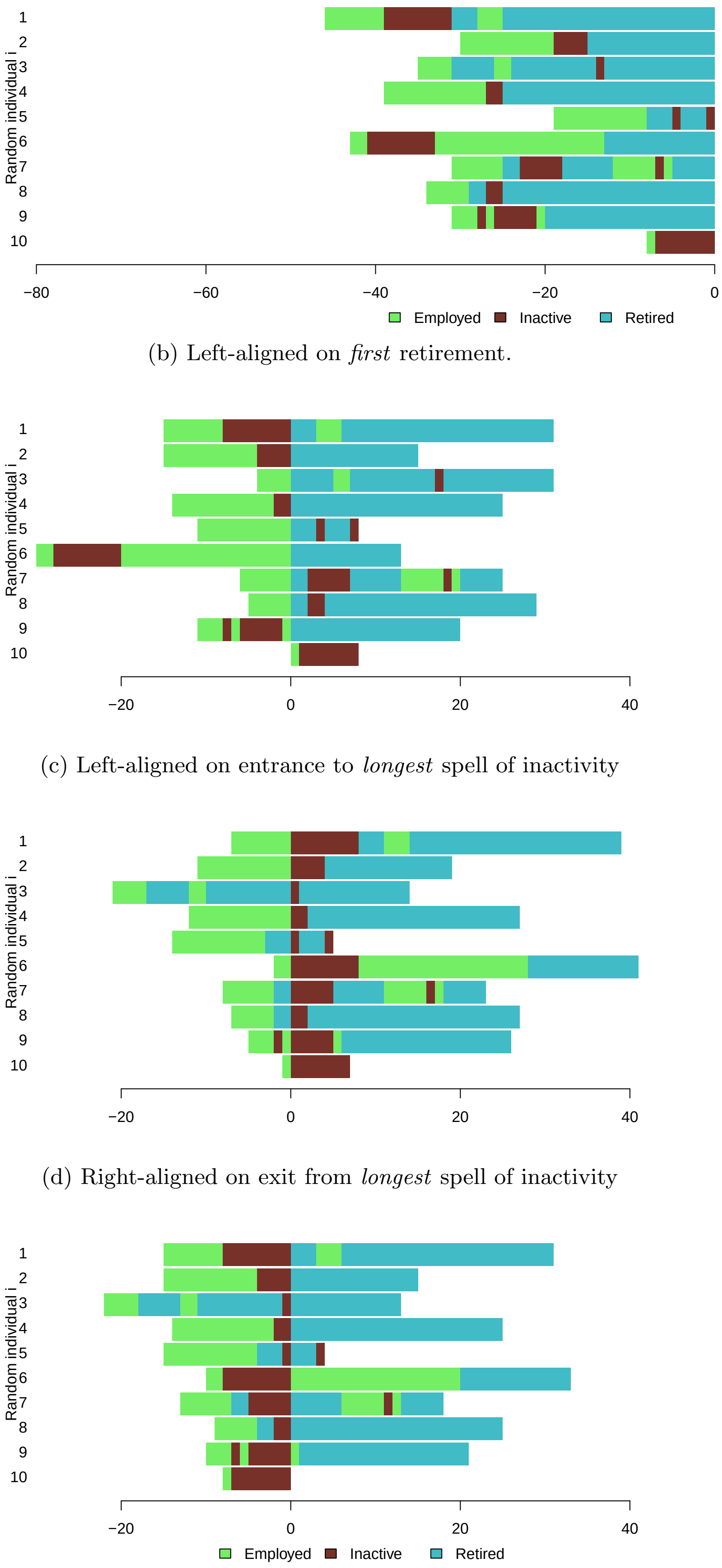
For late-life processes, birth is usually decades away from the events and states of interest, and sharper empirical regularity may be be found with respect to other alignment criteria. Aligning lifelines requires two choices: 1) a reference moment or anchoring *event* must be selected, and 2) the alignment direction must be chosen. A reference event could be any instance of entry, exit, or other compelling anchor point, such as a spell midpoint– ergo such events may relate to episodes themselves. For repeated events, the choice of anchoring episode could itself follow a regular criterion, such as first, last, or longest episode. The *direction* of alignment could be left, right, center, or perhaps something else.

Fig. 5 shows a set of four alignment selections out of the many possible choices. Fig. 5b left-aligns on entry to *first* retirement (if any). One could also choose last, longest, or some other episode of retirement, or of course right-align on exit. Fig. 5c left-aligns on entry into each individual's longest spell of inactivity, whereas Fig. 5d right-aligns on exit from the same spell.

These examples are subset of many possible alignments, in this case column shifting within rows of a matrix. Alignment as shown here is probably insufficient to reveal patterns if one is visualizing raw trajectories, as in these demonstrative figures. One would probably want to define *sort* operations (row-swapping) for this, and that is not something we have ventured to do at this point [3]. Other visualization techniques (ignoring clocks for now) might follow an alignment operation (e.g. Fasang and Liao 2014). In the present, we instead aggregate up to macro patterns.

[3] Possibly the `TraMineR` universe already has *sort* functionality, we need to check.

Figure 5: The sequences from Figure 1 under a variety of alignment types.



(a) Right-aligned on death.

(b) Left-aligned on *first* retirement.

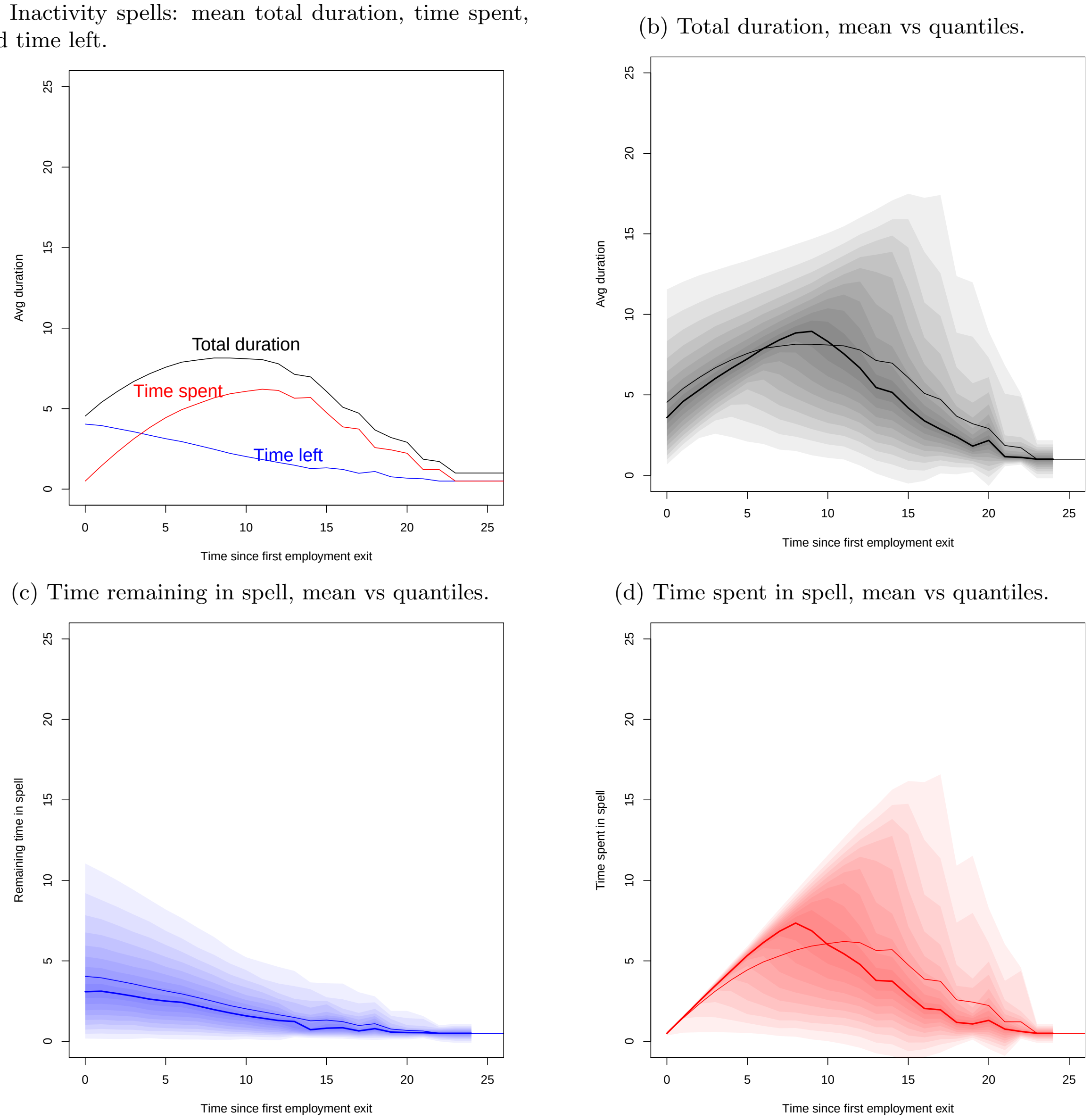
(c) Left-aligned on entrance to *longest* spell of inactivity

(d) Right-aligned on exit from *longest* spell of inactivity

4 Aggregate patterns

Given the choices in clock measures and alignments, the researcher has many degrees of freedom in calculating episode statistics in the aggregate. As a first example, and with no substantive justification as of yet[4], say we'd like to know about inactivity spell patterns by time since first employment exit. We calculate from the same simulated object used for previous exposition. Fig. 6 displays mean conditional episode durations of inactivity structured by time since exiting one's first employment spell, ergo right-aligned on first employment spell and conditional on i) having exited employment, and ii) being in an inactivity spell. Time spent (red, per Fig. 3b) and time left (blue, per Figure 3c) sum to total duration (black, per Fig. 3a) as one would hope. Figs 6b, 6c, and 6d show that mean statistics deviate from median and don't necessarily represent the underlying distribution for any of these three measures.

Figure 6: Inactivity spell statistics by time since end of first employment. Bold lines are median.



(a) Inactivity spells: mean total duration, time spent, and time left.

(b) Total duration, mean vs quantiles.

(c) Time remaining in spell, mean vs quantiles.

(d) Time spent in spell, mean vs quantiles.

[4] We'll swap these examples out with something more substantively compelling.

5 Discussion

We propose a formal set of data operations to allow for creative derivation of demographic macro patterns. Our examples makeup a small subset of those possible, even with the small state space used in our example. To give a sense of the number of macro patterns possible, multiply (1) the number of state categories, (2) episode selection options (first, last, longest, etc), (3) alignment options (left, right, center, etc), and clock options (duration, time spent/left, order, and many others), and it becomes evident that we might have produced over one hundred different macro patterns, just for this relatively simple example case.

It may be surprising to notice that most of these patterns, even though ours resulted from a simple Markov process with a small set of simple and monotonic age patterns, have some character to them. They contain information. Presumably the age patterns that entered into said Markov model do not capture the entire story, and raw observed state sequences are expected to bear stronger degrees of co-dependency. And if we wish to learn something new about a given demographic process, the researcher has (i) large degrees of freedom in selecting macro episode patterns, and (ii) is limited only by one's own creativity in doing so.

The purpose of episode clocks and sequence realignment is to detect important patterns in data (or model results) that are likely to otherwise go unnoticed. Some reasonable priors might include that (i) life course events condition each other; (ii) temporal proximity to life course transitions is likely to be an important predictor of other transitions; (iii) within-episode patterns of other characteristics might be monotonically increasing or decreasing, concave, or convex. Aggregate patterns derived after such operations may be sharper and of more obvious interpretation and consequence than are age patterns.

Promises

This manuscript is an early draft. In a future version we will offer vignette-style applications for a selection of different demographic phenomena, including health and fertility/family demography, with a variety of programmatically generated macro patterns.

References

Hal Caswell and Virginia Zarulli. Matrix methods in health demography: a new approach to the stochastic analysis of healthy longevity and dalys. *Population health metrics*, 16(1):8, 2018.

C. Dudel and M. Myrskylä. Estimating the expected number and length of episodes using markov chains with rewards. (under review), 2017a.

Christian Dudel and Mikko Myrskylä. Working life expectancy at age 50 in the united states and the impact of the great recession. *Demography*, Oct 2017b. ISSN 1533-7790. doi: 10.1007/s13524-017-0619-6. URL `https://doi.org/10.1007/s13524-017-0619-6`.

Anette Eva Fasang and Tim Futing Liao. Visualizing sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods & Research*, 43(4):643–676, 2014.

Simona Iacobelli and Bendix Carstensen. Multiple time scales in multi-state models. *Statistics in medicine*, 32(30):5315–5327, 2013.

Cecilia Potente and Christiaan Monden. Disability pathways preceding death in england by socio-economic status. *Population studies*, 72(2):175–190, 2018.

Marcel Raab, Anette Fasang, and Moritz Hess. Pathways to death: The co-occurrence of physical and mental health in the last years of life. *Demographic Research*, 38:1619–1634, 2018.

Tim Riffe, Pil H Chung, Jeroen Spijker, and John MacInnes. Time-to-death patterns in markers of age and dependency. *Vienna Yearbook of Population Research*, 14:229–254, 2016.

Tim Riffe, Jonas Schöley, and Francisco Villavicencio. A unified framework of demographic time. *Genus*, 73(1):7, 2017.

Giorgio Alfredo Spedicato. Discrete time markov chains with r. *The R Journal*, 07 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-036/index.html`. R package version 0.6.9.7.