Alignment, clocking, and macro patterns of episodes in the life course
(a very rough first draft)

Tim Riffe[*1]

[1]Max Planck Institute for Demographic Research

February 9, 2018

Abstract

Individuals are often modeled as passing through a series of discrete states. Multistate markov models are a means to calculate expectancies and moment statistics of state occupancies under the assumption that the transition rate schedules between states are held fixed. There are however no handy analytic expressions for any but the most simple characteristics of state episodes. Here I demonstrate how to use simulations to estimate a variety of age and other macro patterns of different state episodes. I introduce a suite of episode "clock" measures that count within or between episodes, and the notion of sequence "alignment" as a means to create new structuring axes within which to aggregate clock measures. Resulting aggregate patterns can be age patterns of episode duration, or a variety of other potential measures. This sequence aggregation toolkit may also be used to pre-process sequences prior to a standard sequence analysis, so as to yield more apt clusters.

1 Introduction

The amount of population-level measures that one might devise for a particular process is dizzying, so it may beg the question from the outset why we might desire to have more. Incidence-based matrix models are rather undeveloped with respect to tenure-statistics, and these might be of interest for a variety of substantive reasons. By tenure statistics I refer to the statistics of particular spells or episodes of a state. For example, in incidence-based models with bidirectional flows (i.e., allowing for recovery and then re-onset, and so on), modelled individuals may pass through a state, such as sickness, many times before death. Typically a transition matrix manipulation would only give us the average time spent sick or moment statistics thereof. Recently matrix calculations have been described for how to calculate the average number of episodes of a given state (Dudel and Myrskylä 2017a). Combined with the average state occupancy, this information yields the average duration of episodes.

One may wonder how the average spell duration changes with age, and for this there is no ready matrix expression. Such analytic expressions are surely possible, but in the following I intend to propose a suite of operations (including the aforementioned) so broad and flexible as to fill the lifetime of a matrix buff. I proceed using simulations rather than matrix calculations because it will save the work of deriving and checking dozens of formulas. In this way, we have the liberty to change definitions without incurring methodological setbacks. Since we simulate, we get stochastic stationary distributions of each measurement for free, which I'll represent using fan-chart visualizations. This approach is not all that different from that proposed by Laditka and Wolf (1998), but I expand on their approach by proposing a set of count statistics called *clocks* and age *realignments*, which together result in a large (unbounded?) set of age-like patterns of state episodes. These same operations can be used on observed populations of sequences as well, but the key difference is that observed populations are not stationary. Our example will refer to the stationary tenure statistics that belong to a particular set of transition rates.

---

[*]riffe@demogr.mpg.de

2   Data

I use a published transition matrix from a recent study of working life expectancy in the United States (Dudel and Myrskylä 2017b) as the basis of my example stationary population. This matrix refers to black females aged 50-100 in 1994. The same following steps can be done with any age-stage matrix. One can indeed also do the same with an ageless transition matrix of the life course, if the simulated sequence steps are interpretable as age steps.

3   Simulation

I take advantage of the recently published R package markovchain (Spedicato 2017), which includes a random state sequence generator function rmarkovchain() that merely requires a transition matrix to do its work.[1] I generate a large number of trajectories (10k) to operate on as the stationary population. Each trajectory consists in a character vector of states [Employed,Inactive,Retired,Dead], where dead is of course an absorbing state, but the other states can be switched on and off annually from ages 50 and higher.[2]

   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 done on the full simulated population. In this case, I simulated assuming that one starts in a state of employment at age 50, but the starting state could also easily be a mixture of states.

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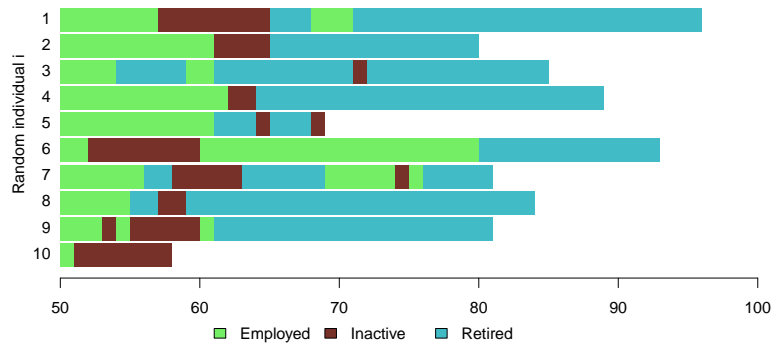

   Standard calculations of prevalence typically proceed by imputing reference states with 1s (with 0s elsewhere) and taking column means over survivors in each age. Figure 2 shows such a data construct, where the state sequence matrix has been converted to a binary matrix, with 1s for employment episodes, 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.

---

[1]There are some trivial object definition steps to convert the employment matrix into a conformable markov object before feeding it to the random generator.

[2]It would also be possible to further graduate transition rates to standardized month or week units to produce higher resolution sequences, but this is unnecessary for the present treatment.

Figure 2: Binary imputation of employment spells
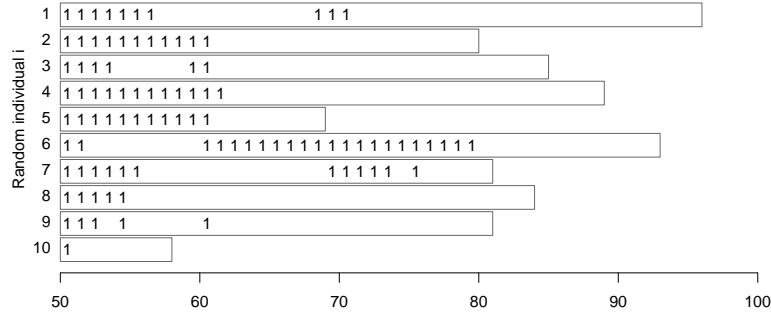
4   Running clocks and alignment

Beyond counting episodes, one may wish to aggregate statistics on each episode in novel ways. For instance, conditional on being in state s in age x, what is the average duration of spell that one finds oneself in? Or time spent or left in the state? Or state order?

4.1   Clocks

To calculate the average spell duration by age, first convert our state sequences to a data object something like Figure 3a. Using chronological age as the reference, one may also wish to calculate time spent or left in the state episode, per Figure 3b or 3c [3]. In either of these cases, value alignment is with respect to episode entry or exit, but aggregation alignment remains pegged to age. Statistics across individuals in an array will therefore produce age patterns.

One may fill episodes with other markers, such as episode order, as in Figure 4 for the case of employment spells, or episode fractions. One could further condition age patterns of total duration, time spent, or time left on episode order. If spells are filled with 1s, then aggregation results in prevalence. Note, time *left* in the episode has no left-truncation problem, also not in the aggregate. These series of values, that I call clocks, are then aggregated in some way. Aggregation proceeds within some external structuring classes defined in the following section.
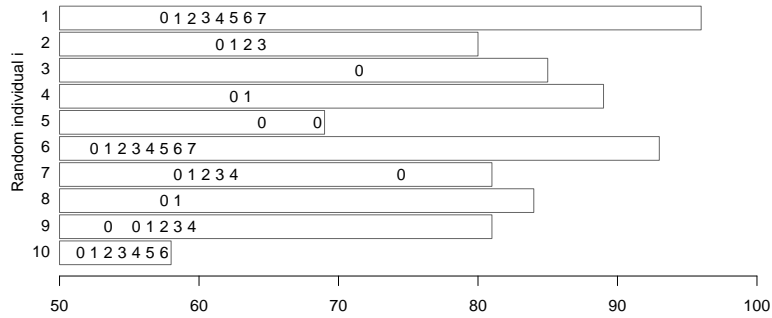
---

[3] Actually, I'd increment values by $\frac{1}{2}$ for mid-state clocking, but decimals would squeeze the figure too much.

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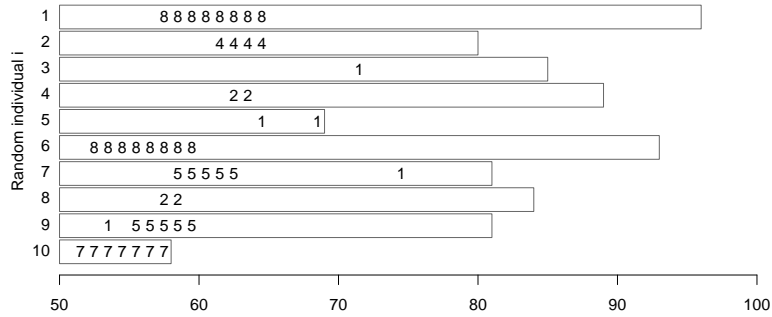
(b) Running; Time spent in episode of inactivity.

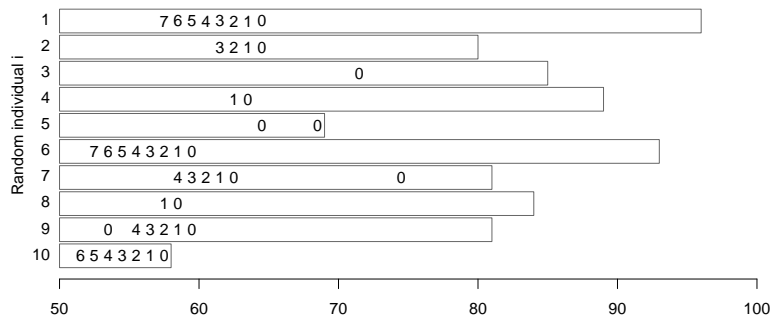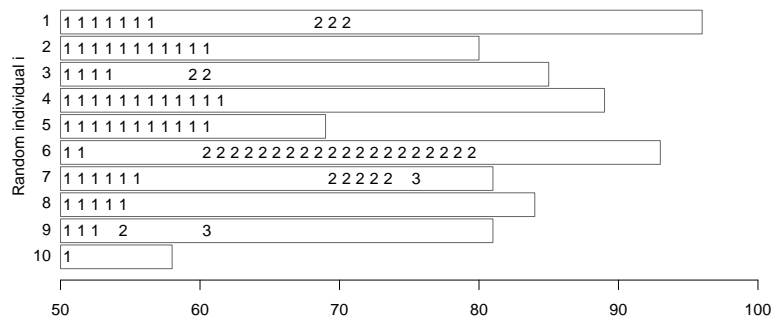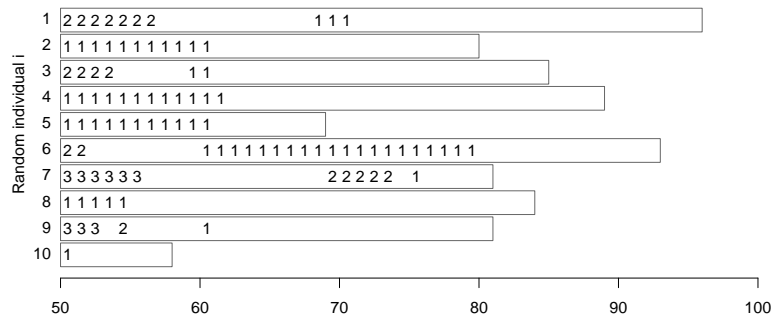(c) Running; Time left in episode of inactivity

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

(a) Employment episode order, increasing.

(b) Employment episode order, decreasing.

4.2 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.

Aggregated patterns would certainly turn out different if we were to right-align on the moment of death, per Figure 5a. This particular realignment doesn't seem so compelling for the demonstrated process, but I suppose it would be illuminating for health states and the sequence of events leading up to death.

Figure 5 shows a set of four alignment selections out of the many possible choices. Figure 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. Figure 5c left-aligns on entry into each individual's longest spell of inactivity, whereas Figure 5d right-aligns on exit from the same spell.

The purpose of realigning is to bring transitions into focus. We do not showcase *sorting* in this treatment, as this is another can of worms, likely in need of sophisticated clustering and sequence analysis. For researchers doing sequence analysis, count and alignment operations would precede standard analysis, and one would need to be judicious in choosing a clustering algorithm that is less sensitive to realignment. For the present, I would rather like to operate on aggregations of individual sequences, in which case between-individual sorting is unimportant. We'd still like to make statements about population-level characteristics.

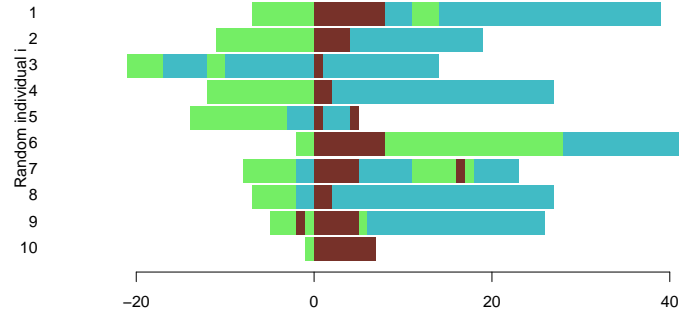Figure 5: The sequences from Figure 1 under a variety of alignment types.

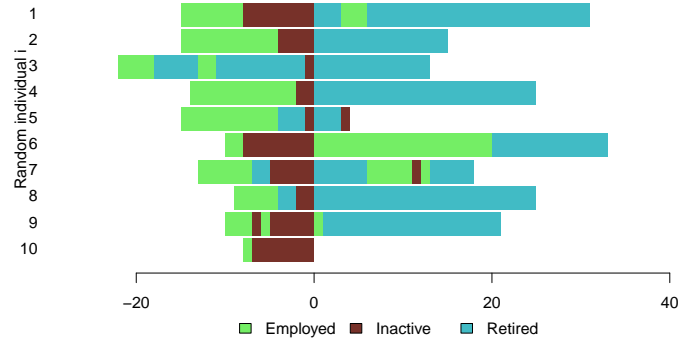(a) Right-aligned on death.
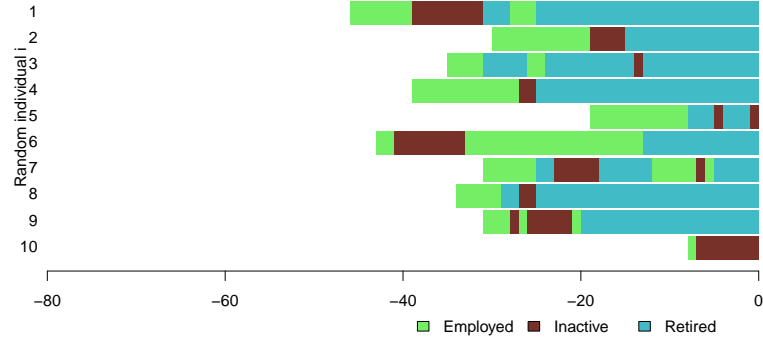
(b) Left-aligned on *first* retirement.
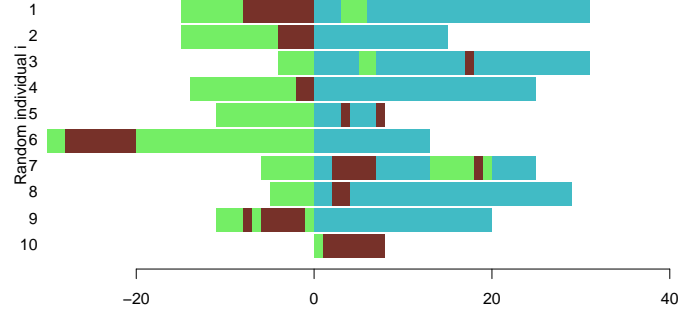
(c) Left-aligned on entrance to *longest* spell of inactivity

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

5 Aggregate patterns

Given a multistate process, and 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, say we'd like to know about inactivity spell patterns by time since first employment exit. Such patterns can be calculated directly on the same simulated object used for previous exposition. Figure 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 Figure 3b) and time left (blue, per Figure 3c) sum to total duration (black, per Figure 3a) as one would hope. Figures 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, while normal lines are mean.
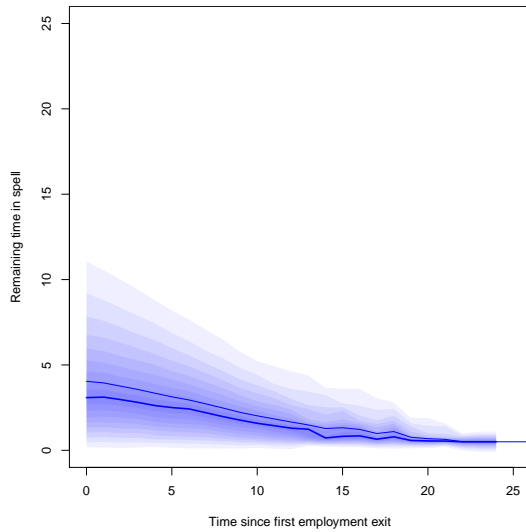
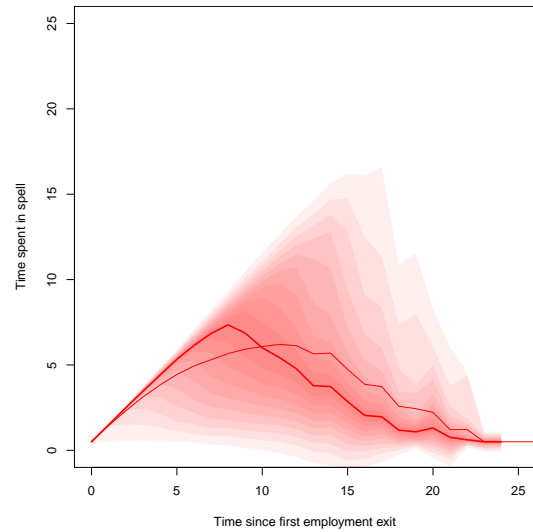(a) Inactivity spells: mean total duration, time spent, and time left.
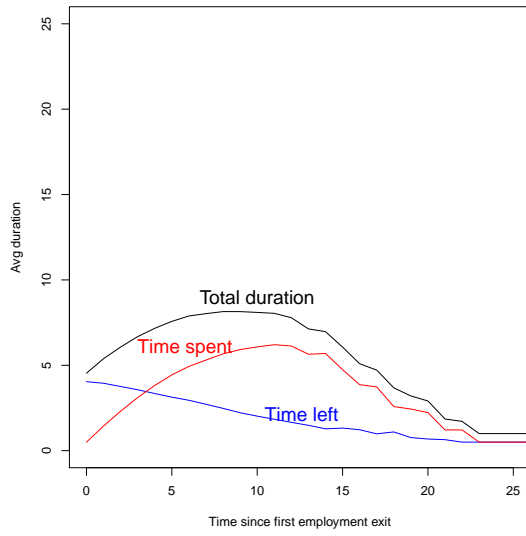
(b) Total duration, mean vs quantiles.
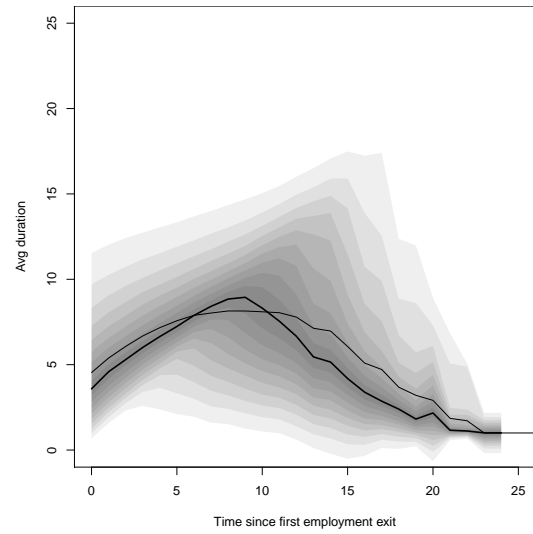
(c) Time remaining in spell, mean vs quantiles.

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

6 Discussion

Observe that the episode statistics displayed in Figure 6, inactivity spells structured by time since exit from one's first employment spell, was chosen haphazardly. Let's say we have a process with 3 state categories. Given one of these states, one may select the first, last, longest or some other episode for purposes of alignment. For simplicity let's say we have 3 choices like this for episode order. Alignment may then be left, right, center, or something else, so let's say again for simplicity that there are 3 ways to align. Now we have 3 or 5 or more clock variables than might be worth calculating statistics on structured like this. It should be clear that the researcher easily has 100+ potential macro age-like patterns that might be calculated, just for this relatively simple example case. It may be surprising to behold that most of these patterns, even though they result 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 codependency. And if we wish to know something about episodes of the states we are studying, 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 researcher is not obliged to move in the direction I propose, but the straw-man default scenario is to treat episode structure (and any process for that matter) as if the only thing that matters is the amount of time passed since birth. Where health, employment, or other life events are co-dependent, the importance of age as a structural predictor is expected to wane with respect to other transitions. Fetuses, babies, and toddlers are largely synchronized with each other as a function of time until or since birth. This is less so for children, whose development slowly begins to lose alignment, far less so for young adults, some of which have passed through first life transitions and others that have passed through none. Adults in mid life and old age are maximally heterogenous with respect to age structure. To gain further insight, it makes sense to realign.

As stated, 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, or concave or convex. Therefore, it may be the case that aggregate patterns derived from such operations are simply sharper and of more obvious interpretation and consequence than are age patterns themselves.

It is difficult to offer advice on the judicious application of data operations proposed here. All I can say is that the researcher should sit back and think creatively about the process under study and not be afraid to experiment some with clocks and alignment. The initial motivating case here involved no re-alignment whatsoever: I just wanted to know how episode duration changes over age. This very question already begs a second round of questions, for it is unclear how to relate episodes to age, given that an age is a point and an episode is a duration. This conundrum is what gives rise to clocks as an operation. From there alignment is a natural next question, and the magnitude of our researcher degrees of freedom increases by another order. In combination, these steps already imply some distancing from markov matrix models, as even cutting edge developments in this methodology are far from reaching the degree of flexibility we propose to have.

References

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.

Sarah B Laditka and Douglas A Wolf. New methods for analyzing active life expectancy. *Journal of Aging and Health*, 10(2):214–241, 1998.

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.