MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Recent developments in discrete-time multistate estimation in Stata

Daniel C. Schneider

German Stata Conference, Hamburg, March 28, 2025

Outline

- Multistate methodology
- The **`dtms`** package: Introduction
- The **`dtms`** package: Planned enhancements

Discrete-Time Multistate Estimation

- Very popular in demography and epidemiology
- Discrete set of states, e.g., employment status, health status
- Discrete-time
  - Evenly spaced (intermittently observed) data => longitudinal surveys
- Under Markov assumption, use
  - **(1)** regression models to
  - **(2)** predict transition probabilities, then
  - **(3)** apply matrix formulas to get outcome statistics

Discrete-Time Multistate Estimation

- Selected outcome statistics
  - `LEXP` – state and life expectancies
  - `LRSK` – lifetime risk
  - `MAFN` – mean age at first entry
  - `EPIS` – number of episodes / number of entries
  - `MDUR` – mean duration of episodes
  - `STAB` – state at absorption
  - `MAAB` – mean age at absorption
  - `MAIS` – mean age in state
  - […]

The dtms Package: Introduction

- Big effort, large package
- For a quick package introduction, see this presentation:
  https://www.stata.com/meeting/germany23/slides/Germany23_Schneider.pdf
- For details and formulas, see:
  - Working papers
    Schneider (2023), DOI:10.4054/MPIDR-WP-2023-041
    Schneider / Myrskyla (2023), DOI:10.4054/MPIDR-WP-2023-042
  - The extensive help files of the package, which include many examples
    Current package installation:

```
. net install dtms, from(https://user.demogr.mpg.de/schneider/stata)
```

The dtms Package: Introduction

- `dtms` estimation proceeds in sequential steps:

  (2) regression (`mlogit`)

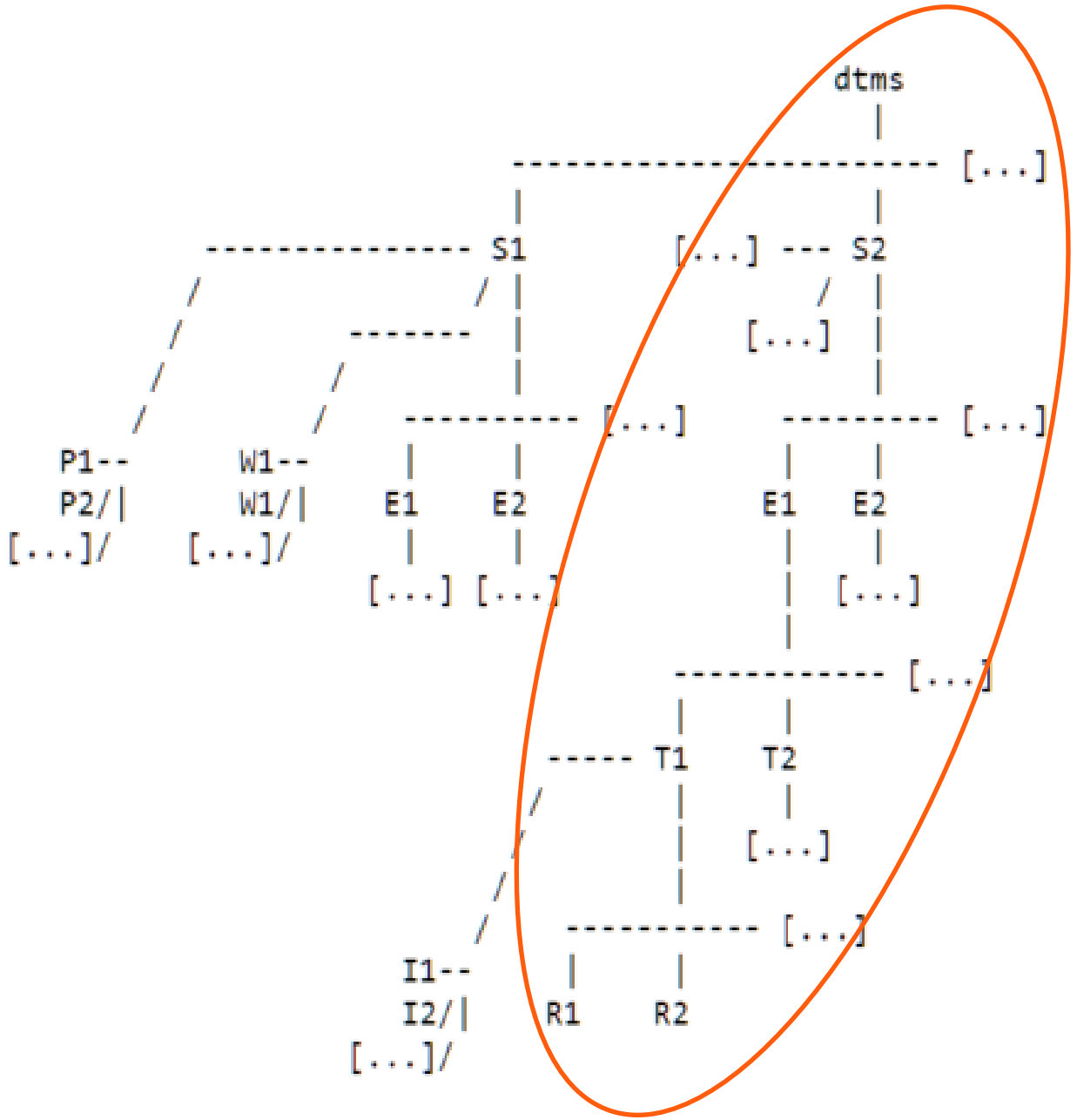
  (3) predict transition probabilities

  (4) calculate various results from them

  Very first step: **(1)** Model setup, so there are a total of four steps

- The results of all steps and their interconnections are organized in the **dtms tree**.

Tree Management



LEVEL 1
MAIN : S-ETUP

- - - - - - - - -

LEVEL 2
SIDE : P-ROPORTION
RE-W-ARDS
MAIN : E-STIMATE

- - - - - - - - -

LEVEL 3
MAIN : T-RANSPROB

- - - - - - - - -

LEVEL 4
SIDE : S-I-MDATA
MAIN : R-ESULT

Tree Management

```
. dtms dir
  (S) impex : (no label) | tra IDs: 1 2 3 | abs IDs: 4 | 61 ages: 50-110
    (P) pfixed : (no label) | fixed:  0.880 0.100 0.020
    (E) mlog : (no label) | cmdline: mlogit cog3 iL.cog3 c.age c.age#c.age se..
      (T) allmeans : (no label) | dtms trans atmeans: L.cog3=(1 2 3) age=(1 2..
        (R) lexp : (no label) | prop: pfixed | timing: mid | calc: analytic |..
      (T) edlow : (no label) | dtms trans atmeans: L.cog3=(1 2 3) age=(1 2 3 ..
        (R) lexp : (no label) | prop: pfixed | timing: mid | calc: analytic |..
        (R) stab : (no label) | prop: pfixed | timing:  | calc: analytic | ST..
      (T) edhigh : (no label) | dtms trans atmeans: L.cog3=(1 2 3) age=(1 2 3..
        (R) lexp : (no label) | prop: pfixed | timing: mid | calc: analytic |..
        (R) stab : (no label) | prop: pfixed | timing:  | calc: analytic | ST..
```

```
. dtms file save using temptree.dtms , replace
```

Example Result

```
. dtms result stab (s1 e1 t1 r1) , initpr(pfixed) form(%7.2f) post nopv
```

State distribution before absorption

state	init none	mild	sevr	total
none	0.60	0.55	0.36	0.59
mild	0.29	0.33	0.21	0.30
sevr	0.11	0.12	0.43	0.11
total	1.00	1.00	1.00	1.00

State distribution before absorption:

		Coefficient	Std. err.	[95% conf.	interval]
none					
	none	0.598	0.027	0.545	0.650
	mild	0.295	0.023	0.250	0.340
	sevr	0.107	0.019	0.070	0.144
mild					
	none	0.552	0.026	0.500	0.604
	mild	0.332	0.023	0.286	0.378
	sevr	0.116	0.020	0.077	0.156
sevr					
	none	0.358	0.033	0.293	0.422
	mild	0.212	0.024	0.165	0.258
	sevr	0.431	0.048	0.336	0.525
cond					
	none	1.000	0.000	1.000	1.000
	mild	1.000	0.000	1.000	1.000
	sevr	1.000	0.000	1.000	1.000
state					
	none	0.588	0.026	0.537	0.640
	mild	0.297	0.023	0.253	0.341
	sevr	0.115	0.019	0.078	0.151
total					
	total	1.000	0.000	1.000	1.000

Package Highlights / Contributions

- A **framework for organizing** and saving results
- Up to 10 states, exactly one of which is absorbing
- Automated estimation of transition probabilities **(speed)**
- **14+ different outcome statistics**
- **Asymptotic covariance matrices**
- **Partial age ranges**
- **Group comparisons**
  linear and nonlinear hypothesis testing on any number of and on any type of results
- **Speed** (interactive time range)
- **Markov chains with rewards**
  general implementation; includes the possibility of user-defined rewards
- Works with `svy` estimation
- Generation of data sets with simulated trajectories
- Simulation-based results as alternative to analytical ones

Main Package Enhancements

- The following changes are currently being implemented but not yet publicly available.

Main Package Enhancements

- Incorporate initial state distribution into asymptotics

State distribution before absorption

	init			
	none	mild	sevr	total
state				
none	0.60	0.55	0.36	0.59
mild	0.29	0.33	0.21	0.30
sevr	0.11	0.12	0.43	0.11
total	1.00	1.00	1.00	1.00

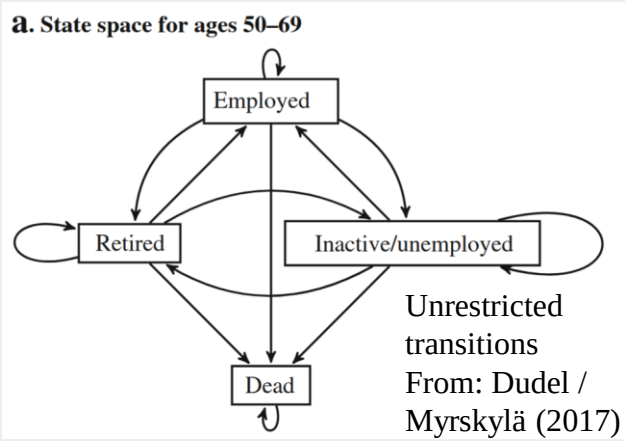
State distribution before absorption:

	Coefficient	Std. err.	[95% conf.	interval]
state				
none	0.588	0.026	0.537	0.640
mild	0.297	0.023	0.253	0.341
sevr	0.115	0.019	0.078	0.151
total				
total	1.000	0.000	1.000	1.000

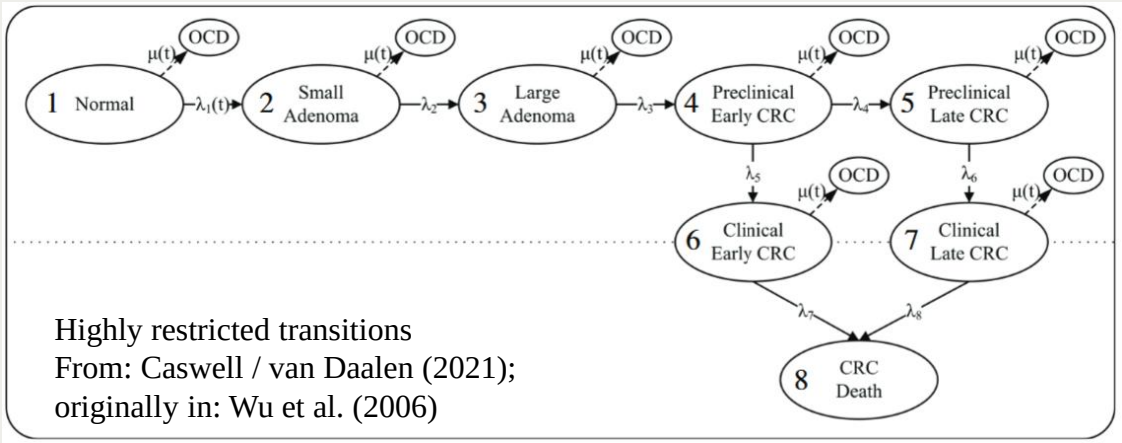
- Treat initial population proportions estimates as uncorrelated ?
- Make use of stable prevalences (Brouard 2019) ?
- Stable distribution of transition matrix at first model age ?

Main Package Enhancements

- Restricted transitions



Unrestricted transitions
From: Dudel / Myrskylä (2017)

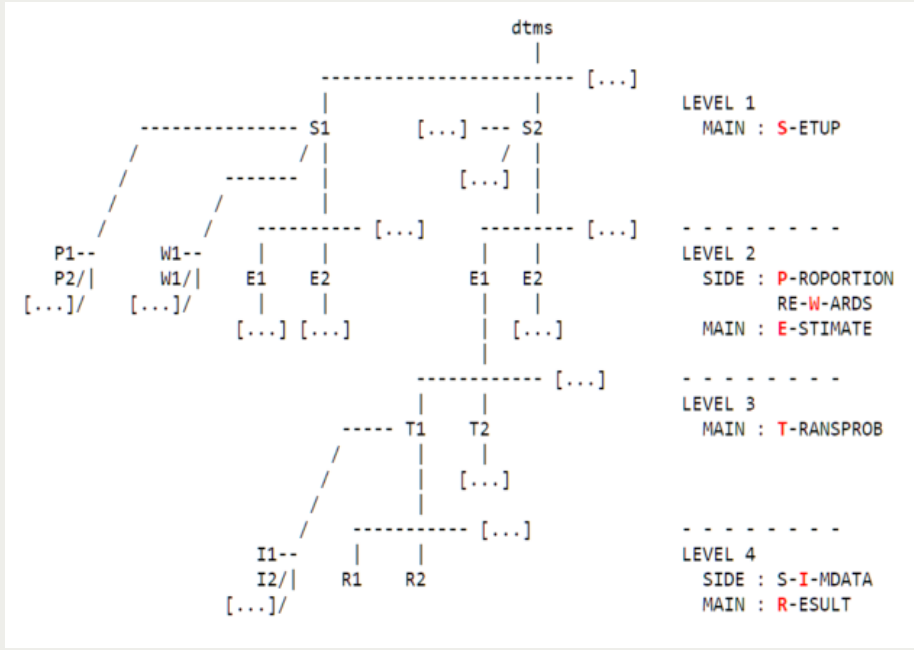


Highly restricted transitions
From: Caswell / van Daalen (2021);
originally in: Wu et al. (2006)

- For few restrictions, `mlogit` is already working
- For many restrictions, convergence problems or inefficiencies
- Potential solutions:
  - Restrict `mlogit` coefficients?
  - Estimate individual `logits` and then combine?

Main Package Enhancements

- Reloading saved dtms files across package versions
  - Difficult problem: Mata class changes
  - Solution (imperfect):
    - Include older Mata libraries (mlib files) in current package version
    - New command `dtmsversion` activates old libraries, which then read trees saved by older versions
    - No direct mechanism of converting older trees to newer ones
    - Maybe: new command `dtms dofile` creates a tree replication script



Main Package Enhancements

- Transition probabilities based on time-varying prediction values
  - This has many applications
  - Examples:
    - Varying an exogenous variable at a certain age
    - Age dummies (e.g., for retirement age)
    - Age splines

Main Package Enhancements

- Import of IMaCh regression results
  - IMaCh short for “Interpolated Markov Chain"
  - Available as a standalone executable (https://euroreves.ined.fr/imach/)
  - Accommodates irregularly spaced data and relaxes the assumption of no unobserved transitions
  - Estimates a multinomial logistic model at a time step frequency that is higher than that of the data source
  - Has a fixed set of results tables and graphs
  - The import of the estimated IMaCh mlogit results into `dtms` unlocks the full set of `dtms` results for users of IMaCh

Main Package Enhancements

- Import of external transition probabilities (matrix P) and their covariance matrix (matrix cov[P])
  - Various use cases, one of them is:
- Link to continuous-time (CT) models

  Idea:
  - Estimate CT model externally
  - Externally generate P for a very fine time grid, along with cov[P]
  - Import P / cov[P] into `dtms`
  - Generate `dtms` results
  - Caveat: no group comparisons will be possible, as this necessitates cov[P1 P2], which (I believe) no current CTMS software can generate

Thank you

schneider@demogr.mpg.de

References

Brouard, Nicolas (2019): Theory and Applications of Backward Probabilities and Prevalences in Cross-Longitudinal Surveys. In Handbook of Statistics (Bd. 40, S. 435–486). Elsevier. DOI:10.1016/bs.host.2018.11.009

Brouard, Nicolas (2021): Computing Health Expectancies Using IMaCh: A maximum likelihood computer program using interpolation of Markov chains. Paris: Institut National d'Etudes Demographiques (INED) and EUROREVES. Available at: euroreves.ined.fr/imach/.

Caswell, Hal, and Silke van Daalen (2021): Healthy Longevity from Incidence-Based Models: More Kinds of Health than Stars in the Sky. Demographic Research 45: 397–452. DOI:10.4054/DemRes.2021.45.13.

Dudel, Christian, and Mikko Myrskylä (2017): Working Life Expectancy at Age 50 in the United States and the Impact of the Great Recession. Demography 54 (6): 2101–23. DOI:10.1007/s13524-017-0619-6.

Schneider, Daniel C. (2023): Statistical Inference for Discrete-Time Multistate Models: Asymptotic Covariance Matrices, Partial Age Ranges, and Group Contrasts." MPIDR Working Paper WP-2023-041. DOI:10.4054/MPIDR-WP-2023-041.

Schneider, Daniel C., and Mikko Myrskylä (2023): Statistical Inference for Discrete-Time Multistate Models: Extensions to Markov Chains with Rewards. MPIDR Working Paper WP-2023-042. DOI:10.4054/MPIDR-WP-2023-042.

Wu, G., Wang, Y.M., Yen, A., Wong, J.M., Lai, H.C., Warwick, J., and Chen, T. (2006). Cost-effectiveness analysis of colorectal cancer screening with stool DNA testing in intermediate-incidence countries. BMC Cancer 6 (136): 1–12. DOI:10.1186/1471-24076-136.