Post-Secondary Enrollment in the United States: Model Validation and Student Life Tables

Diana Gonzalez

California State Polytechnic University, Pomona

David Spencer

University of North Carolina, Chapel Hill

Faculty Advisor

Ariel Cintron-Arias

East Tennessee State University

Funding provided by the National Science Foundation, award DMS-1263009.

DEPARTMENT *of*
MATHEMATICS & STATISTICS
College of Arts & Sciences
EAST TENNESSEE STATE UNIVERSITY

Datasets

- Datasets were retrieved from archives managed by:
  - National Center for Education Statistics (https://nces.ed.gov).
  - Integrated Postsecondary Education Data System (https://nces.ed.gov/ipeds/).
- Fall enrollments were used.
- Enrollment totals include:
  - Undergraduate students of all ages.
  - Full-time students.
- Years from 1980 to 1998.
- Institutions:
  - Public universities participating in Title IV of the Higher Education Act.
  - Bachelor's degree granting.

Terminology of Matrix Population Models

- **School departures** (similar to death in demography)
  - Retention rate: it measures the percentage of first-time students who are seeking bachelor's degree who return to the institution to continue their studies the following fall. Similar to survival probability.
  - Drop out rate: it can be approximated by scaling the graduation rates and retention rate.
  - Graduation rate: it tracks the progress of students who began their studies as full-time, first-time degree-seeking to see if they complete a degree or other award such as certificate within 150% of "normal time" for completing the program in which they are enrolled.
  - Causes of departure:
    - Student death
    - Transferring out
    - Leave of absence
    - Graduation
    - Dropout
- **School arrivals**
  - Recruitment into postsecondary school.
  - Similar to birth.
  - Similar to immigration.

Matrix Population Models

- **Classification**
  - Freshman
  - Sophomore
  - Junior
  - Senior
- The state variable denotes the size of each student classification at a time point.
- Specifically,
  - $x_1(n)$ denotes the number of freshman at year n.
  - $x_2(n)$ denotes the number of sophomore at year n.
  - $x_3(n)$ denotes the number of junior at year n.
  - $x_4(n)$ denotes the number of senior at year n.
- The general form of a matrix model is

$$x(n+1) = Ax(n) + b(n)$$

- The matrix *A* stores probabilities.
- The vector *b* is usually called an immigration or recruitment vector.

Matrix Population Models

Model 1

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & p_3 & 0 \end{bmatrix} \times \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} \beta_1 n + \alpha_1 \\ \beta_2 n + \alpha_2 \\ \beta_3 n + \alpha_3 \\ \beta_4 n + \alpha_4 \end{bmatrix}$$

Model 2

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} q_1 & 0 & 0 & 0 \\ p_1 & q_2 & 0 & 0 \\ 0 & p_2 & q_3 & 0 \\ 0 & 0 & p_3 & q_4 \end{bmatrix} \times \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} u(n) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Model 3

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & p_3 & q \end{bmatrix} \times \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} u(n) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Model 4

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & p_3 & q \end{bmatrix} \times \begin{bmatrix} x_1(n-1) \\ x_2(n-1) \\ x_3(n-1) \\ x_4(n-1) \end{bmatrix} + \begin{bmatrix} \beta_1 n + \alpha_1 \\ \beta_2 n + \alpha_2 \\ \beta_3 n + \alpha_3 \\ \beta_4 n + \alpha_4 \end{bmatrix}$$

Ordinary Least Squares with Uncorrelated Observation Errors

$$Y_j = f(t_j, \theta_0) + \mathcal{E}_j$$

- For each time series, we assume the observation is the model output (evaluated at a true parameter) plus random deviations.
- The measurement errors $\mathcal{E}_j$ are assumed:
  - To be independent and identically distributed.
  - To have zero mean.
  - To have constant variance.

Ordinary Least Squares (OLS) with Uncorrelated Observation Errors

- Let $\{y_1, \ldots, y_N\}$ be a realization of the observation process.
- The OLS estimate is

$$\hat{\theta} = \arg\min_{\theta} \sum_{j=1}^{N} (y_j - f(t_j, \theta))^2$$

Model Selection

- Akaike information criterion (AIC) is quantity used to rank mathematical models, according to goodness of fit, number of model parameters, and number of observations.
- Model 2 is best for each of 51 datasets.
- Model 2 has by far the best average.

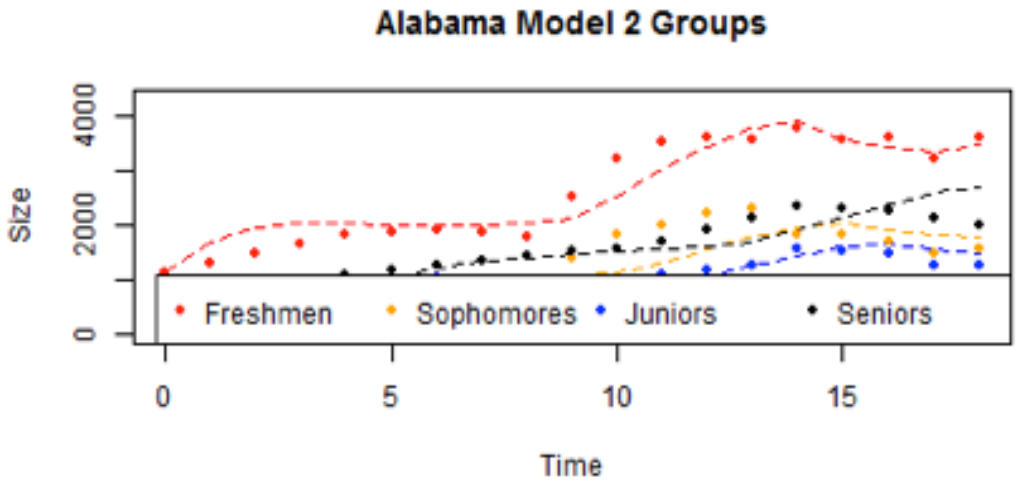
Model	Model 1	Model 2	Model 3	Model 4
Average AIC	***349.1***	***290.9***	***341.7***	***354.6***

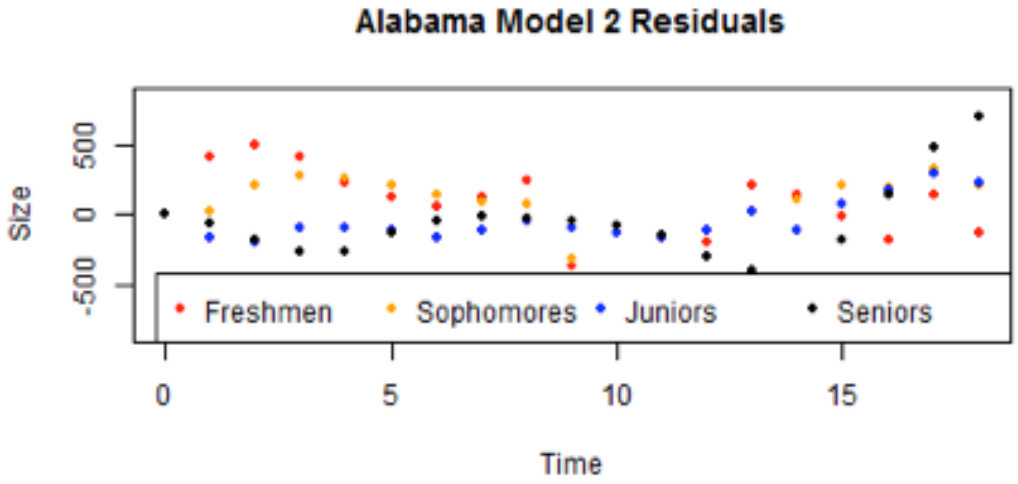
Best Fit Plots and Residual Plots for Model 2

For illustration purposes results for only five (out of fifty one) datasets are displayed:

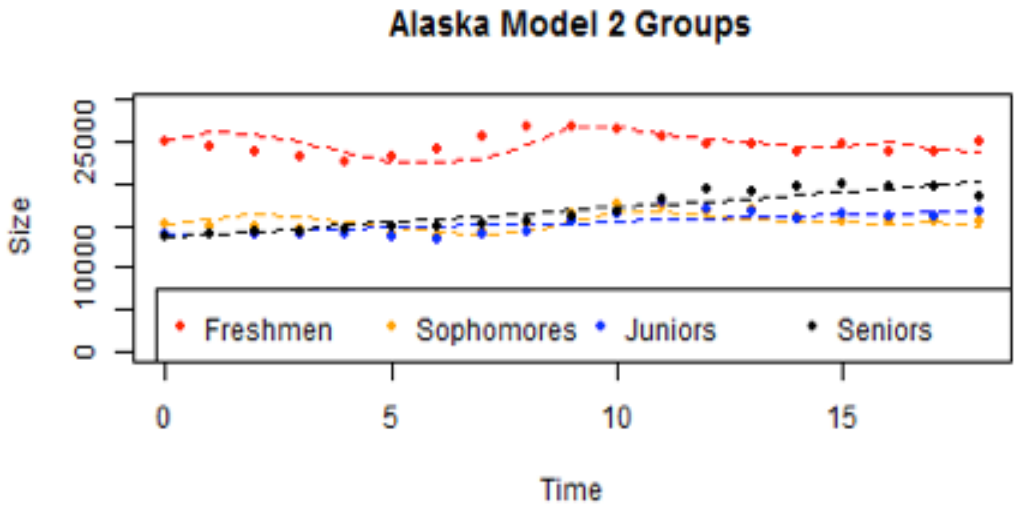
- Alabama
- Alaska
- Arizona
- California
- Colorado

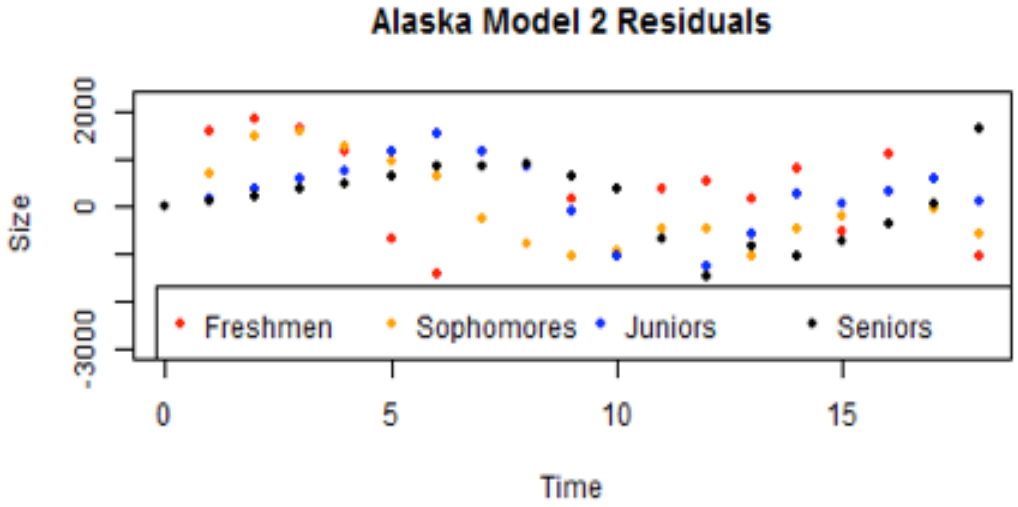
Alabama


Alabama Model 2 Groups
Size
0
2000
4000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

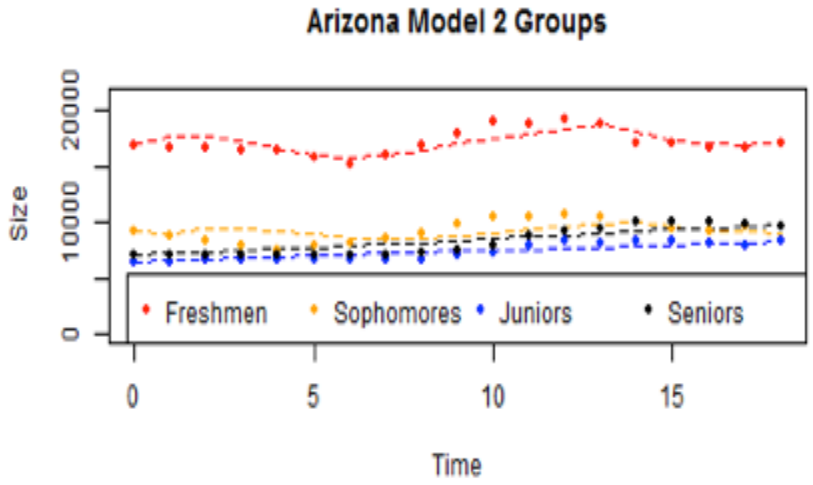

Alabama Model 2 Residuals
Size
-500
0
500
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

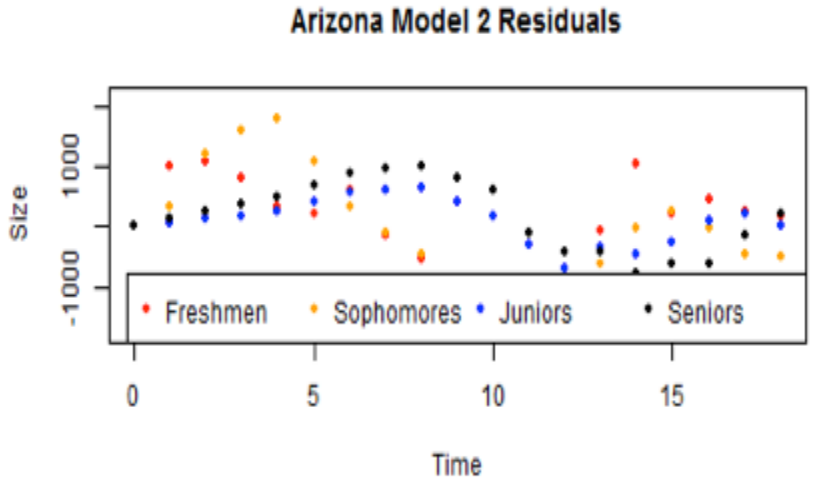
Alaska


Alaska Model 2 Groups
Size
0
10000
25000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

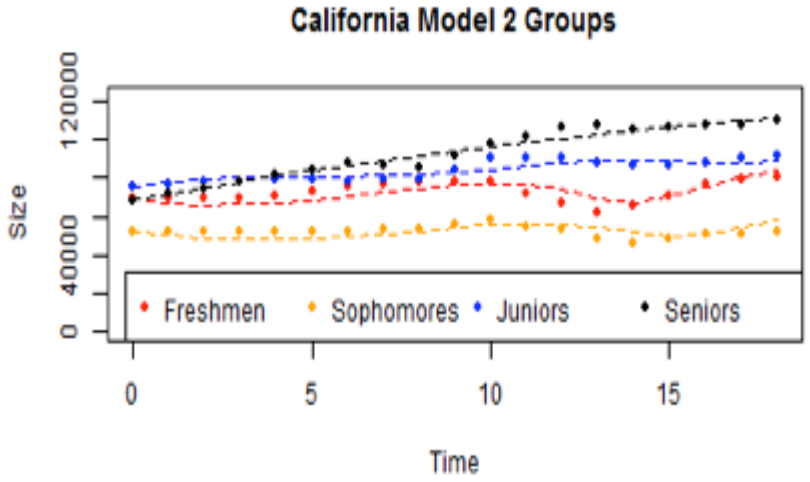

Alaska Model 2 Residuals
Size
-3000
0
2000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

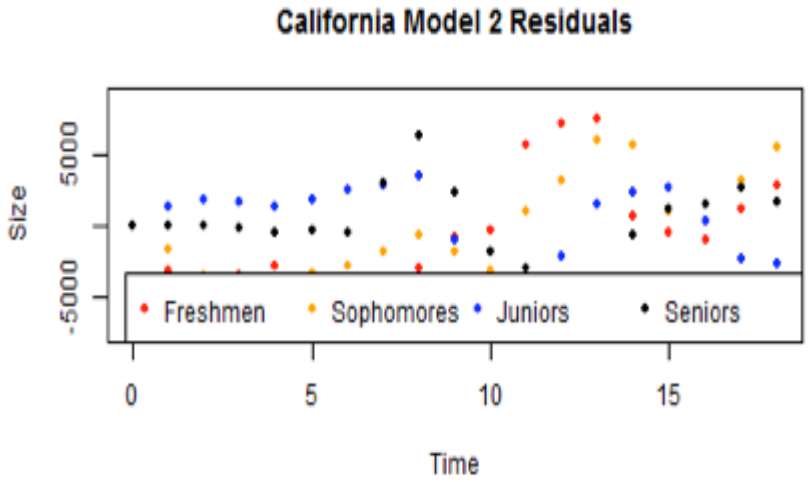
Arizona


Arizona Model 2 Groups
Size
0
10000
20000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

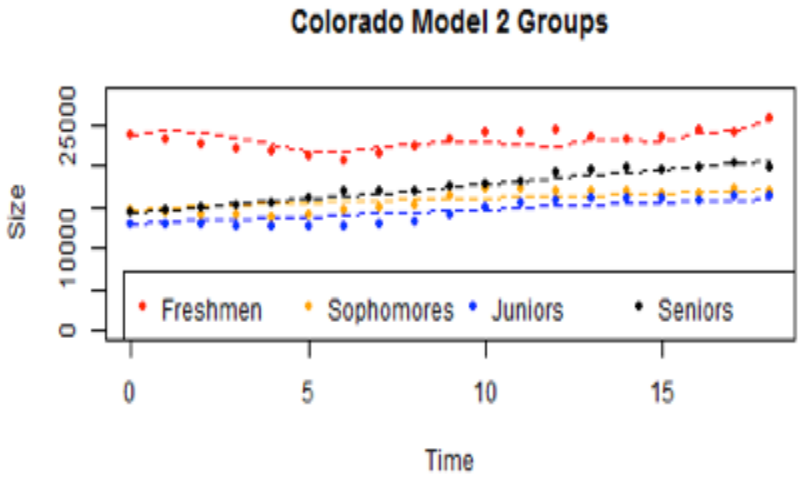

Arizona Model 2 Residuals
Size
-1000
1000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

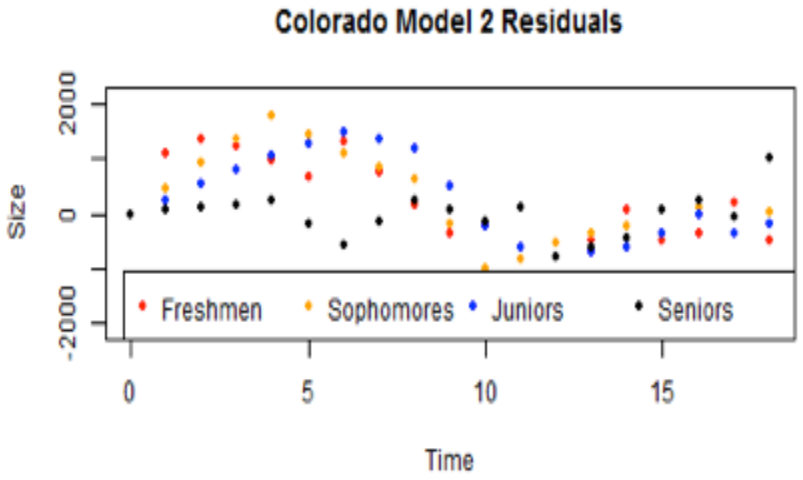
California


California Model 2 Groups
Size
0
40000
120000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time


California Model 2 Residuals
Size
-5000
5000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

Colorado


Colorado Model 2 Groups
Size
0
10000
25000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time


Colorado Model 2 Residuals
Size
-2000
0
2000
Freshmen
Sophomores
Juniors
Seniors
0
5
10
15
Time

Durbin-Watson Test Results

- Durbin-Watson is a test statistic used to determine autocorrelation in the residuals.
- This test can help identify when the *i.i.d*. assumption of the OLS framework is not well-suited.

Student classification	Datasets with autocorrelated residuals
First year	***All states but Arkansas, Hawaii, Idaho, Kansas, Minnesota, Nevada, New Hampshire, New Jersey, Tennessee, Texas, Utah, Vermont, Virginia, and Wvomina***
Second year	***All states but Idaho***
Third year	***All states***
Fourth year	***All states***

Ordinary Least Squares with Correlated Observation Errors

- Suppose the statistical model for the observation process is given by

$$Y_j = f(t_j, \theta_0) + \mathcal{E}_j, \qquad j = 1, \ldots, N,$$

where the observation errors are correlated according to

$$\mathcal{E}_j = \rho_0 \mathcal{E}_{j-1} + \mathcal{U}_j.$$

It is assumed that $\mathcal{U}_j \sim N(0, \sigma_0^2)$, i.e., these random variables are independent and identically distributed.

- Let $\hat{\theta}$ denote the estimate, i.e., the solution of the minimal sum of squares functional:

$$\hat{\theta} = \arg\min_{\theta} \sum_{j=1}^{N} (y_j - f(t_j, \theta))^2.$$

- Define the standardized residuals as follows:

$$\bar{r}_j = \sqrt{\frac{N}{N - \kappa_\theta}} \left(y_j - f(t_j, \hat{\theta}) \right)$$

- Let $d_j = \bar{r}_j - \bar{r}_{j-1}$ for $j = 2, \ldots, N$.

For a fixed m, define $d_1^{(m)} \sim UNIF(-\alpha\bar{r}_1, \alpha\bar{r}_1)$, where $\alpha > 0$ is a tuning parameter.

- Use simple random sampling with replacement to generate $\{d_2^{(m)}, \ldots, d_N^{(m)}\}$ from $\{d_2, \ldots, d_N\}$. Define the mth bootstrap sample of size N as $\{d_1^{(m)}, d_2^{(m)}, \ldots, d_N^{(m)}\}$.

- An estimate of ρ_0 is denoted as $\hat{\rho}$, which can be computed from the standardized residuals

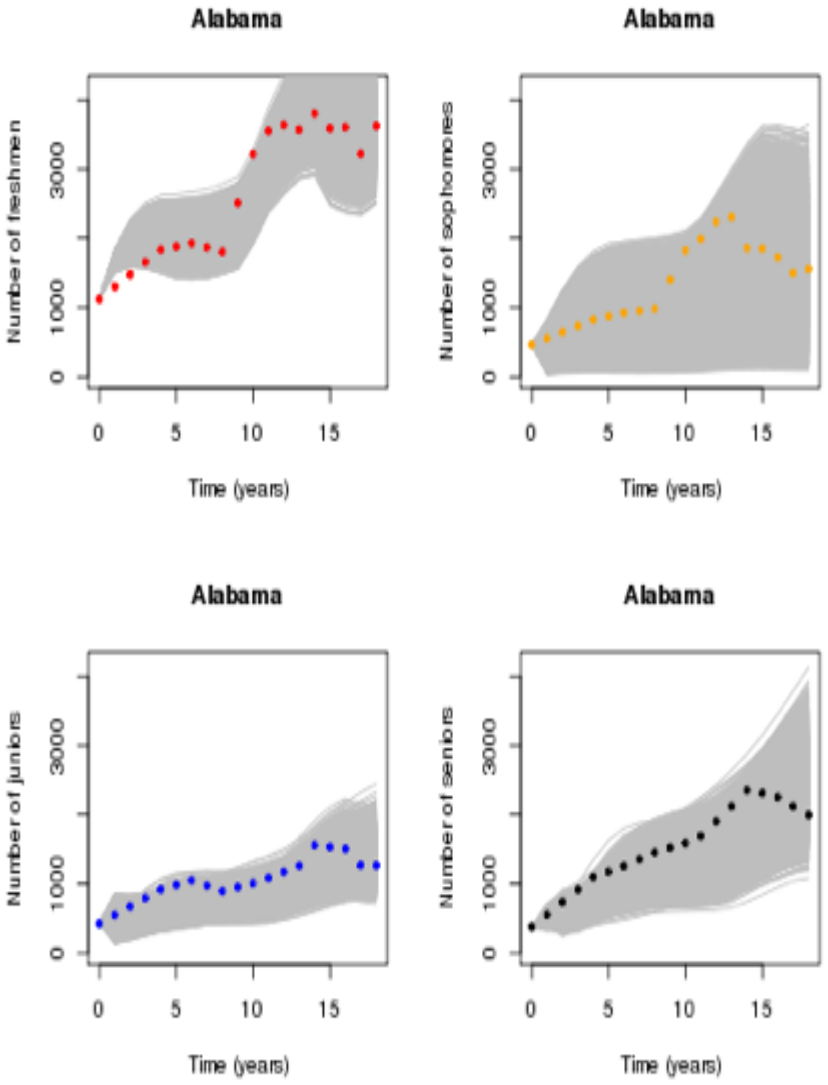
$$\hat{\rho} = \frac{\sum_{j=1}^{N} \bar{r}_j \bar{r}_{j-1}}{\sum_{j=1}^{N} \bar{r}_j^2}$$

Preliminary Results

For illustration purposes results for only five (out of fifty one) datasets are displayed:

- Alabama
- Alaska
- Arizona
- California
- Colorado

Alabama: Uncertainty Clouds from Bootstrap Sampling

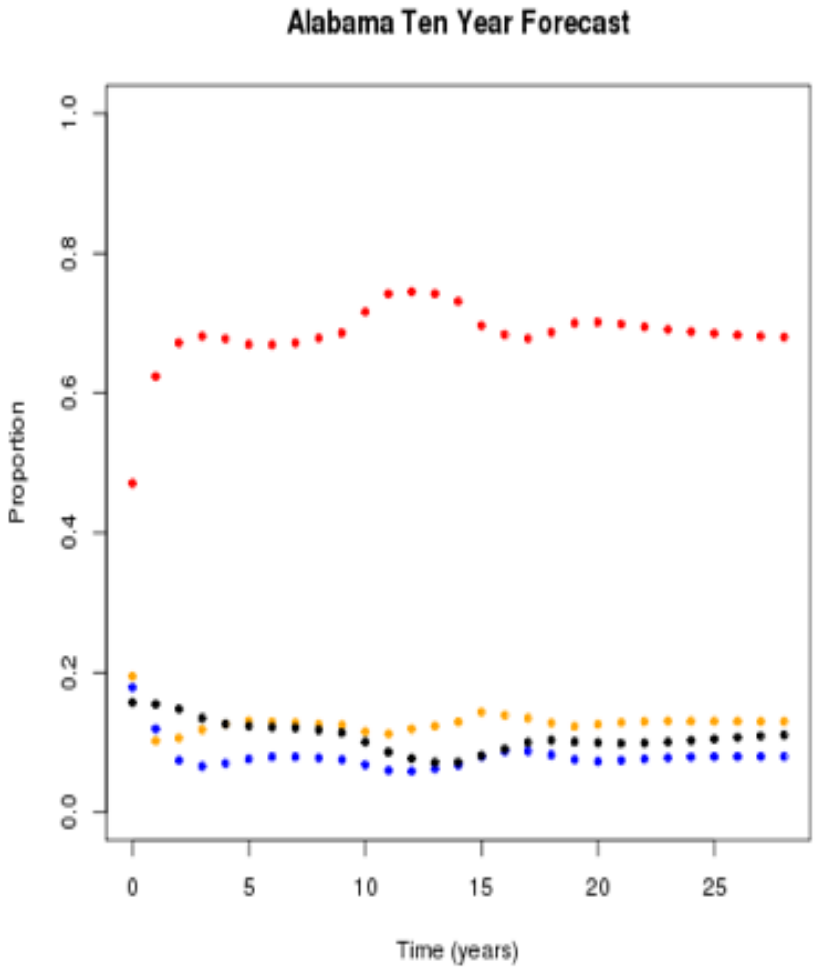


Alabama: Model 2 Parameter Estimates and Standard Errors

Alabama

	p12	p23	p34	p11	p22	p33	p44
Means	0.3179	0.5835	0.4398	0.5611	0.2705	0.3083	0.7939
Standard Errors	0.1246	0.2506	0.326	0.0415	0.2537	0.3157	0.2342

Alabama: Longitudinal Forecast of Proportions

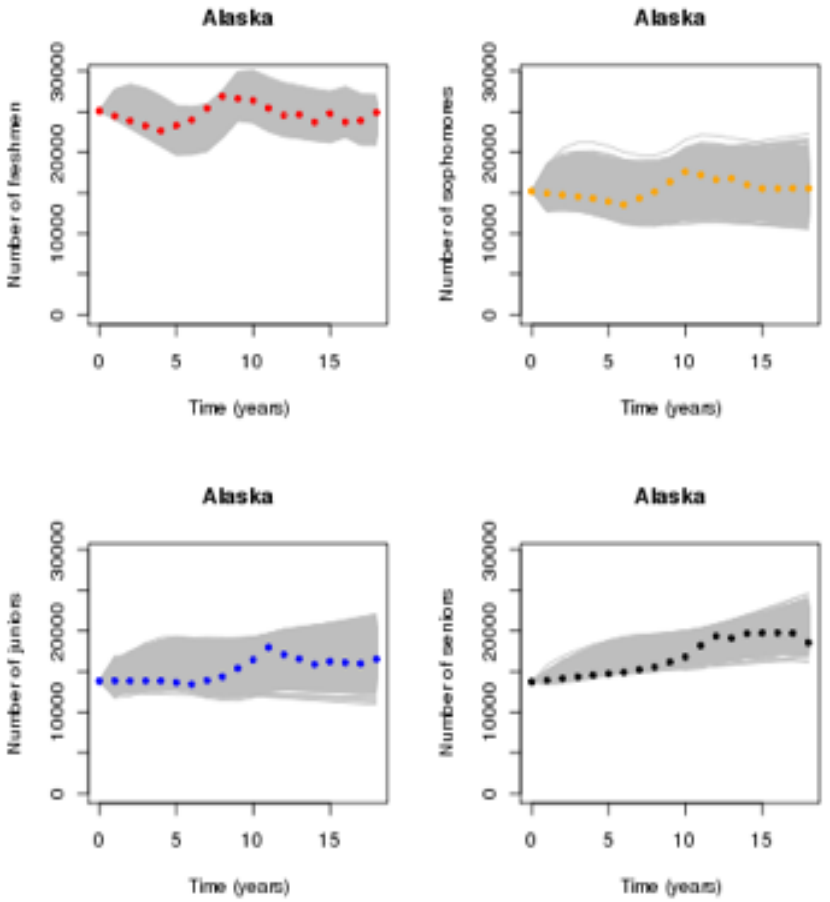


Alabama: Student Life Table

Alabama

	Year 1	Year 2	Year 3	Year 4+
# living in group	10000	7242.0169	5792.5624	3683.0645
Life expectancy	5.8956	4.9948	4.5308	4.852
Year death prob	0.121	0.146	0.2519	0.2061
Group death prob	0.2758	0.2001	0.3642	1
Avg yrs in group	2.2784	1.3708	1.4458	4.852
# dying in group	2757.9831	1449.4545	2109.4978	3683.0645
Yrs lived in group	22784.0655	9927.1157	8374.7542	17870.3814

Alaska: Uncertainty Clouds from Bootstrap Sampling

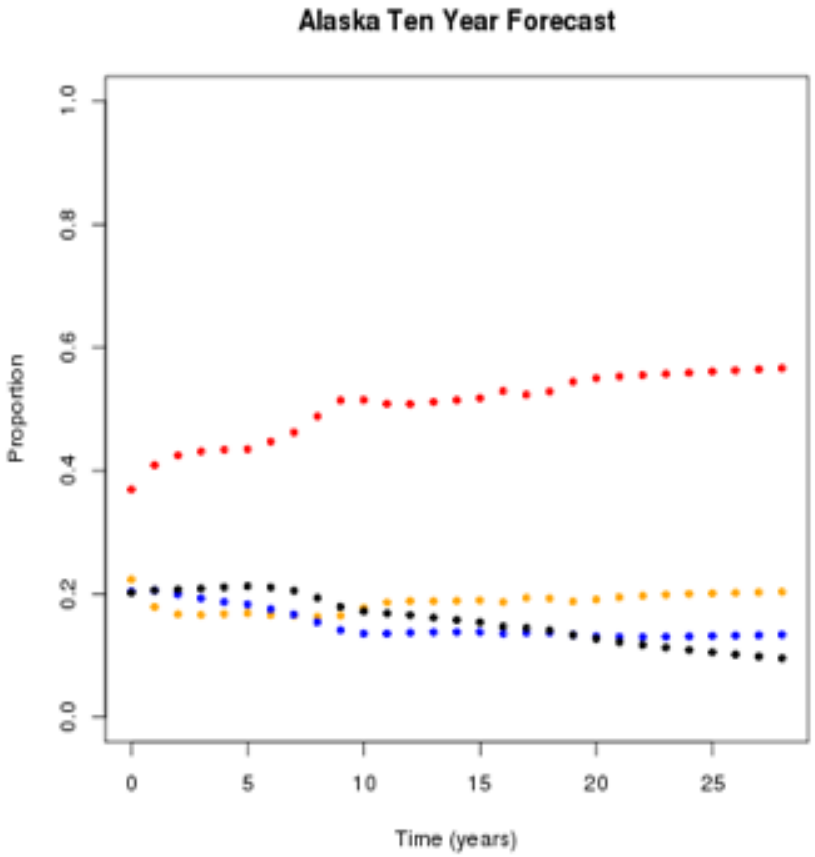


Alaska: Model 2 Parameter Estimates and Standard Errors

Alaska

	p12	p23	p34	p11	p22	p33	p44
Means	0.3751	0.196	0.0814	0.4199	0.3942	0.8212	0.946
Standard Errors	0.1903	0.176	0.0907	0.0204	0.3076	0.1788	0.0864

Alaska: Longitudinal Forecast of Proportions

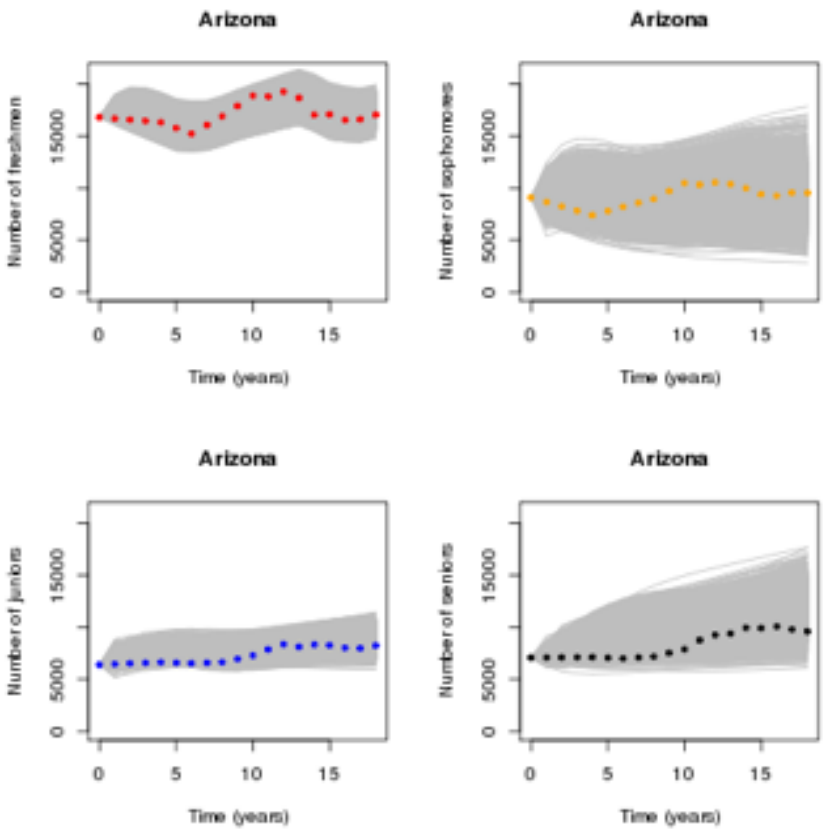


Alaska: Student Life Table

Alaska

	Year 1	Year 2	Year 3	Year 4+
# living in group	10000	6466.6786	2092.0166	952.3607
Life expectancy	5.7246	6.1866	14.0208	18.5115
Year death prob	0.205	0.4098	0.0974	0.054
Group death prob	0.3533	0.6765	0.5448	1
Avg yrs in group	1.724	1.6508	5.5937	18.5115
# dying in group	3533.3214	4374.662	1139.6559	952.3607
Yrs lived in group	17239.7606	10675.0301	11702.0543	17629.5924

Arizona: Uncertainty Clouds from Bootstrap Sampling

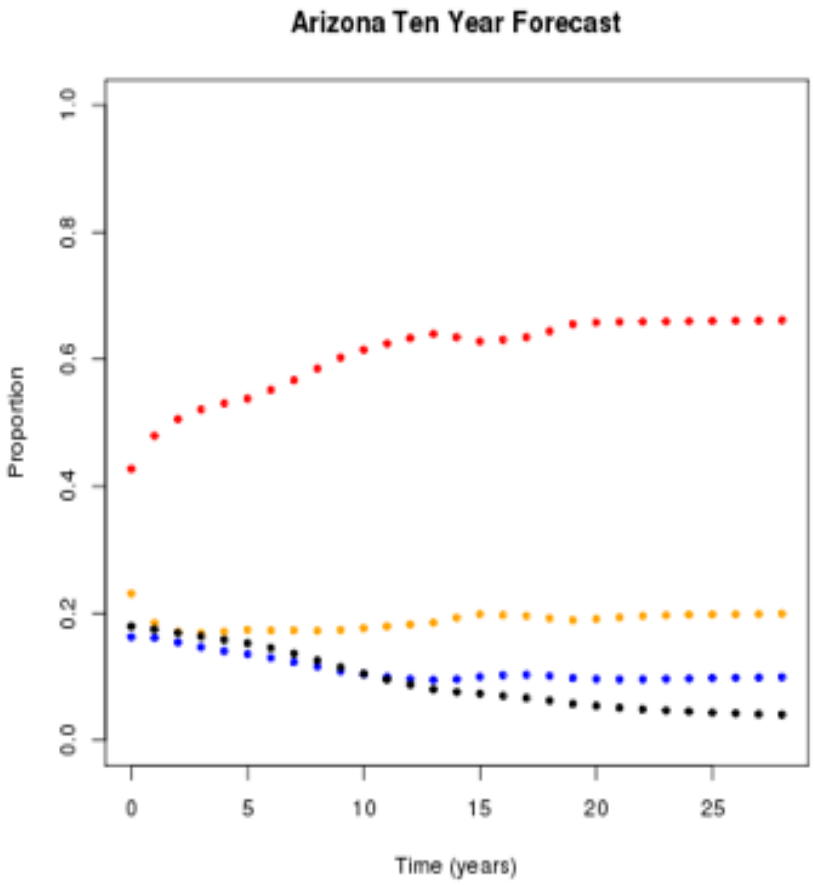


Arizona: Model 2 Parameter Estimates and Standard Errors

Arizona

	p12	p23	p34	p11	p22	p33	p44
Means	0.293	0.1985	0.1771	0.4309	0.4455	0.7818	0.8828
Standard Errors	0.192	0.1609	0.1906	0.0241	0.3544	0.2019	0.1599

Arizona: Longitudinal Forecast of Proportions

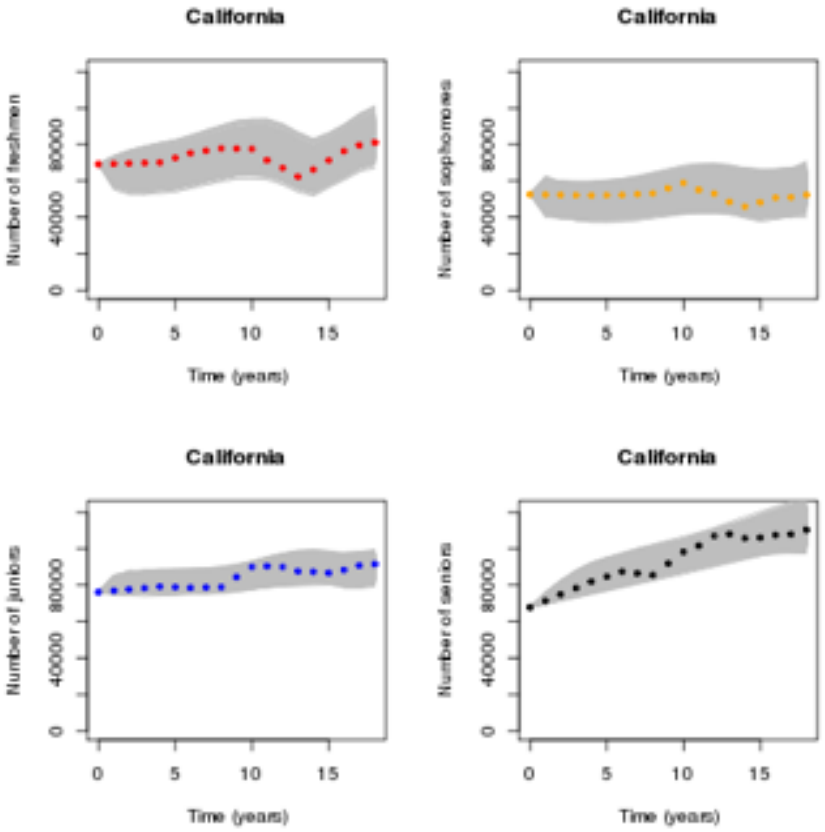


Arizona: Student Life Table

Arizona

	Year 1	Year 2	Year 3	Year 4+
# living in group	10000	5149.7565	1843.7237	1496.2909
Life expectancy	4.8083	5.9245	11.5105	8.5361
Year death prob	0.276	0.356	0.0411	0.1172
Group death prob	0.485	0.642	0.1884	1
Avg yrs in group	1.7573	1.8035	4.583	8.5361
# dying in group	4850.2435	3306.0328	347.4328	1496.2909
Yrs lived in group	17573.0815	9287.4925	8449.7025	12772.4168

California: Uncertainty Clouds from Bootstrap Sampling

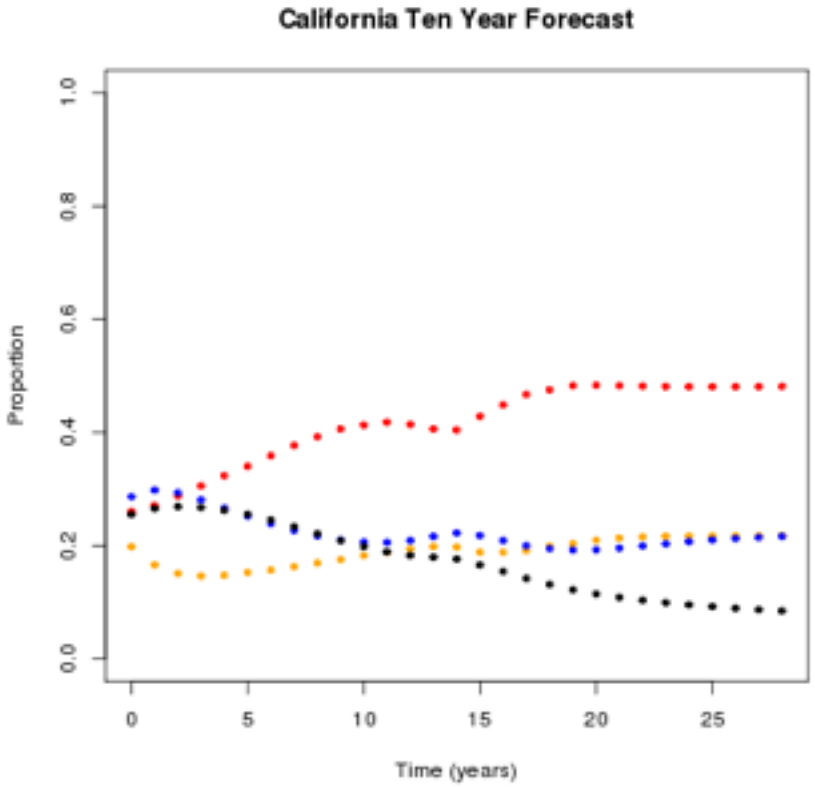


California: Model 2 Parameter Estimates and Standard Errors

California

	p12	p23	p34	p11	p22	p33	p44
Means	0.4442	0.3417	0.1233	0.391	0.4017	0.8005	0.9133
Standard Errors	0.1574	0.1537	0.0675	0.0366	0.2134	0.0965	0.0635

California: Longitudinal Forecast of Proportions

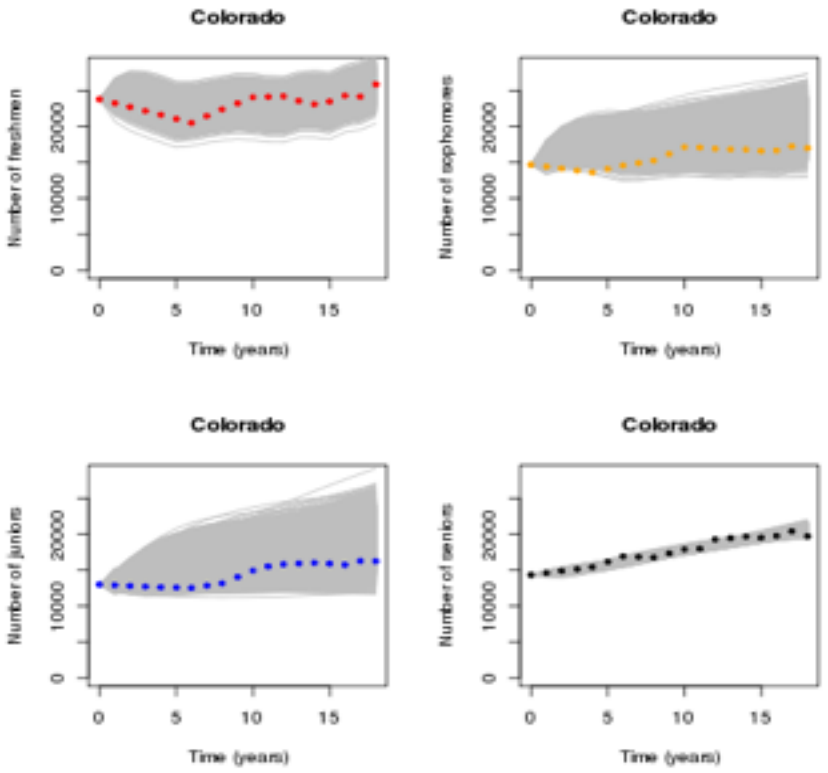


California: Student Life Table

California

	Year 1	Year 2	Year 3	Year 4+
# living in group	10000	7294.2127	4166.212	2574.61
Life expectancy	7.9191	8.6055	12.1402	11.5331
Year death prob	0.1648	0.2566	0.0762	0.0867
Group death prob	0.2706	0.4288	0.382	1
Avg yrs in group	1.642	1.6715	5.013	11.5331
# dying in group	2705.7873	3128.0006	1591.602	2574.61
Yrs lived in group	16420.4812	12191.9712	20885.321	29693.3287

Colorado: Uncertainty Clouds from Bootstrap Sampling

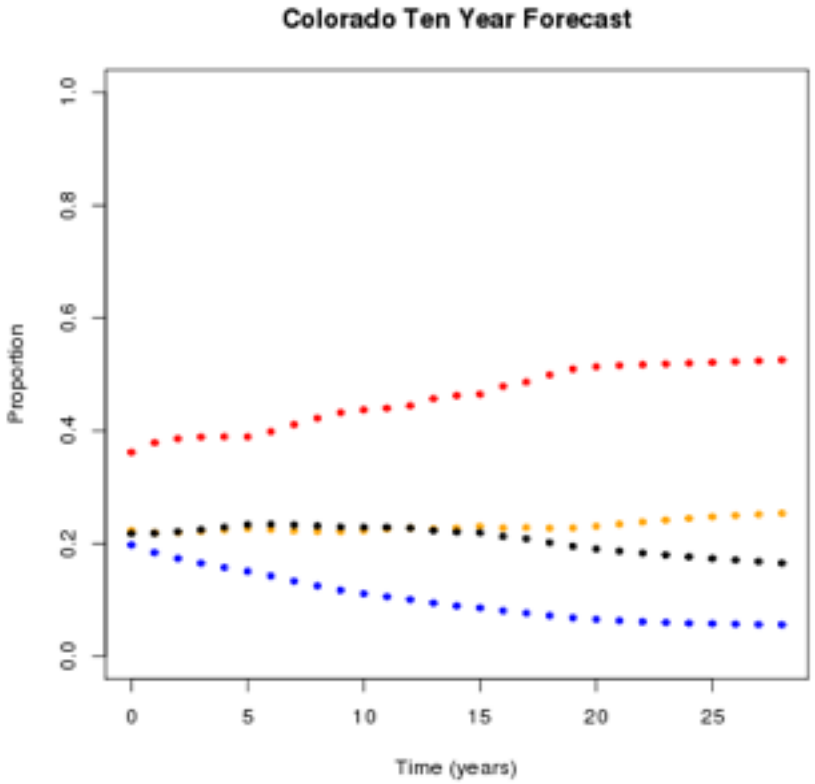


Colorado: Model 2 Parameter Estimates and Standard Errors

Colorado

	p12	p23	p34	p11	p22	p33	p44
Means	0.1487	0.119	0.044	0.4015	0.8193	0.8913	0.9791
Standard Errors	0.1175	0.1173	0.0378	0.0315	0.1658	0.1266	0.0364

Colorado: Longitudinal Forecast of Proportions



Colorado: Student Life Table

Colorado

	Year 1	Year 2	Year 3	Year 4+
# living in group	10000	2484.9131	1636.1616	662.1416
Life expectancy	7.726	24.3679	28.6054	47.951
Year death prob	0.4498	0.0617	0.0647	0.0209
Group death prob	0.7515	0.3416	0.5953	1
Avg yrs in group	1.6708	5.533	9.2001	47.951
# dying in group	7515.0869	848.7515	974.02	662.1416
Yrs lived in group	16707.6249	13748.9619	15052.769	31750.3308