The Visualization of College Enrollment Data

By Collin DeVore

Introduction

- There have been many studies regarding the effects of psychology and culture on the enrollment rates of females and males
- Few studies, however, have attempted to determine if the geographic, political, economic differences, or other such localized processes between states could feasibly create differing processes in the data that are not being captured by the aggregate models, nor have the processes been shown through visualization techniques to be similar or the same
- For this reason, this study uses different calculations and visualizations to show the way in which the ten largest states by population are continuously attempting to close the enrollment gap between males and females
  - This study also examines the possibility that geographic, political, regional, or cultural factors could somehow influence the states to the point that they are not representative of the aggregated dataset

Literature Review

- Charlotte Agger, Judith Meece, and Soo-yong Byun (2018) – Analyzes male and female rural adolescents to observe how perceptions of parental expectations and job opportunities affect each. Finds that males have a higher perception of job opportunities while females have a higher perception of parental expectations.
- Dylan Conger and Mark C. Long (2010) – Examines the reasoning behind why women are earning higher grades, earning more credits, and last longer in college. Finds that males start college with these issues due to leaving high school without having taken care of the issues.
- Claudia Goldin, Lawrence F. Katz, and Ilyana Kuziemko (2006) – Examines the trend in which females began enrolling in college in much larger numbers. Finds that females focus more on preparing for college in high school and that females take advantage of their rights to go to college with the perception of gaining more money.
- Jerry A. Jacobs (1996) – Provides an examination of many studies involving the differences of the opportunities of males and females in the collegiate setting. Mentions that, while females are enrolling more and gaining their degrees more often than in the past, they are still not as common in larger, more prestigious, schools.
- Mark Hugo Lopez and Ana Gonzalez – Barrera (2014) – Article that explained a Pew Research Center study in which the researchers analyzed the fact that women are enrolling more than men even by minority population. The findings show that Hispanic women, in comparison to Hispanic men, are enrolling much more often.

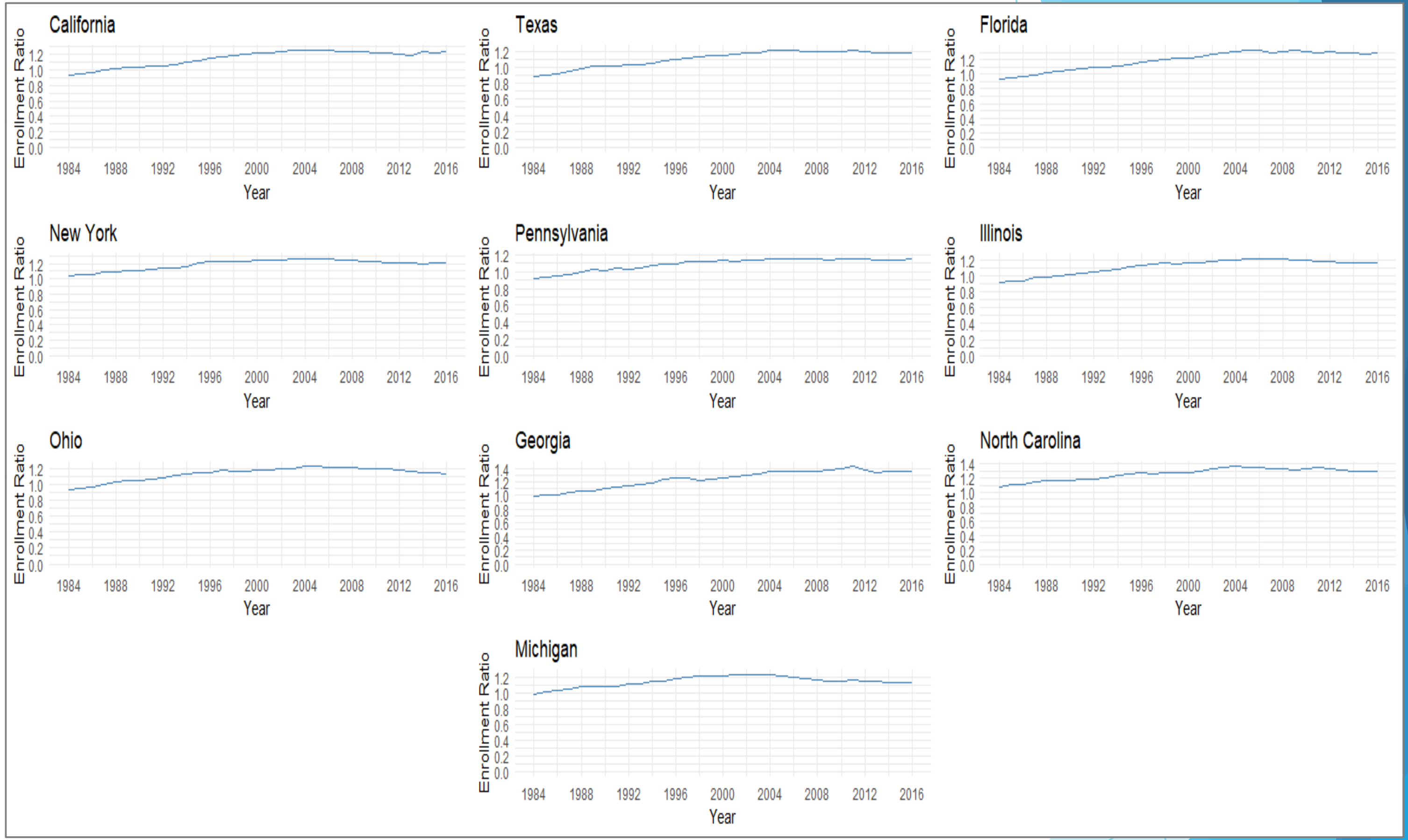
- Each of these studies have analyzed and attempted to understand the reasoning behind the fact that females are enrolling in larger numbers than males

Data

- The data taken for this analysis comes from the Digest of Education Statistics, as reported by the National Center for Education Statistics
- This digest reports on the differing data and statistics within the educational system at a two year lag
- For this analysis, the number of full time fall college enrollments for males and females are taken
  - Only Bachelor's level is considered
- My dependent variable is the enrollment ratio
  - (Number of Females Enrolled)/(Number of Males Enrolled)
- My independent variable is time

1st Visualization of the Data

- These are the graphs for each state of the female/male ratio
- Each of the datasets appear to follow the same general path, trailing off near the end
- Georgia looks as though it is following a straight path, with only a slight curve near the end
- Michigan curves more than the rest of the data, as though it is moving back towards an equal amount of males and females

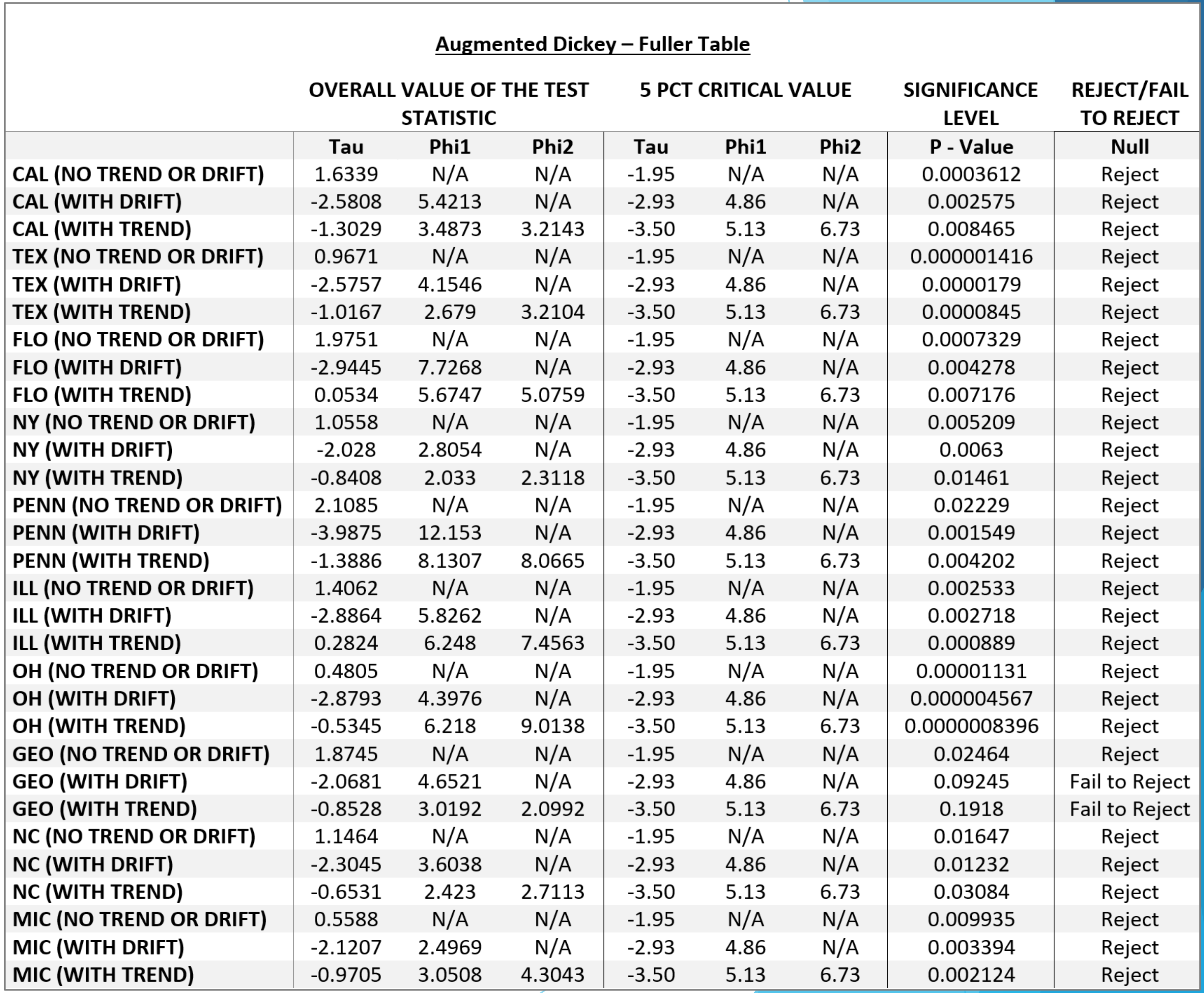


```
library("ggplot2")
g1cal <- ggplot(data = data1, mapping = aes(x = Year, y = California, color = Value)) +
  ggtitle("California") +
  xlab("Year") +
  ylab("Enrollment Ratio") +
  geom_line(stat="identity", color = "steelblue", show.legend = TRUE) +
  theme_minimal() +
  scale_y_continuous(breaks = round(seq(min(0), max(1.45), by = 0.2),1)) +
  scale_x_continuous(breaks = round(seq(min(1984), max(2016), by = 4),1)) +
  expand_limits(x=1984, y=0)
g1cal
```

```
library("gridExtra")
lay1 <- rbind(c(1, 2, 3),
              c(4, 5, 6),
              c(7, 8, 9),
              c(NA, 10, NA))
grid.arrange(g1cal, g1tex, g1flo, g1new, g1pen, g1ill, g1ohi,
g1geo, g1nor, g1mic, layout_matrix = lay1)
```

Augmented Dickey – Fuller Tests

- In order to test stationarity, augmented Dickey – Fuller tests are run
  - The number of lags is chosen by the Akaike Information Criterion
- The decision of whether or not to reject the null hypothesis come from the p – value shown in the graph
- Besides possibly Georgia, none of the data appears to be non stationary

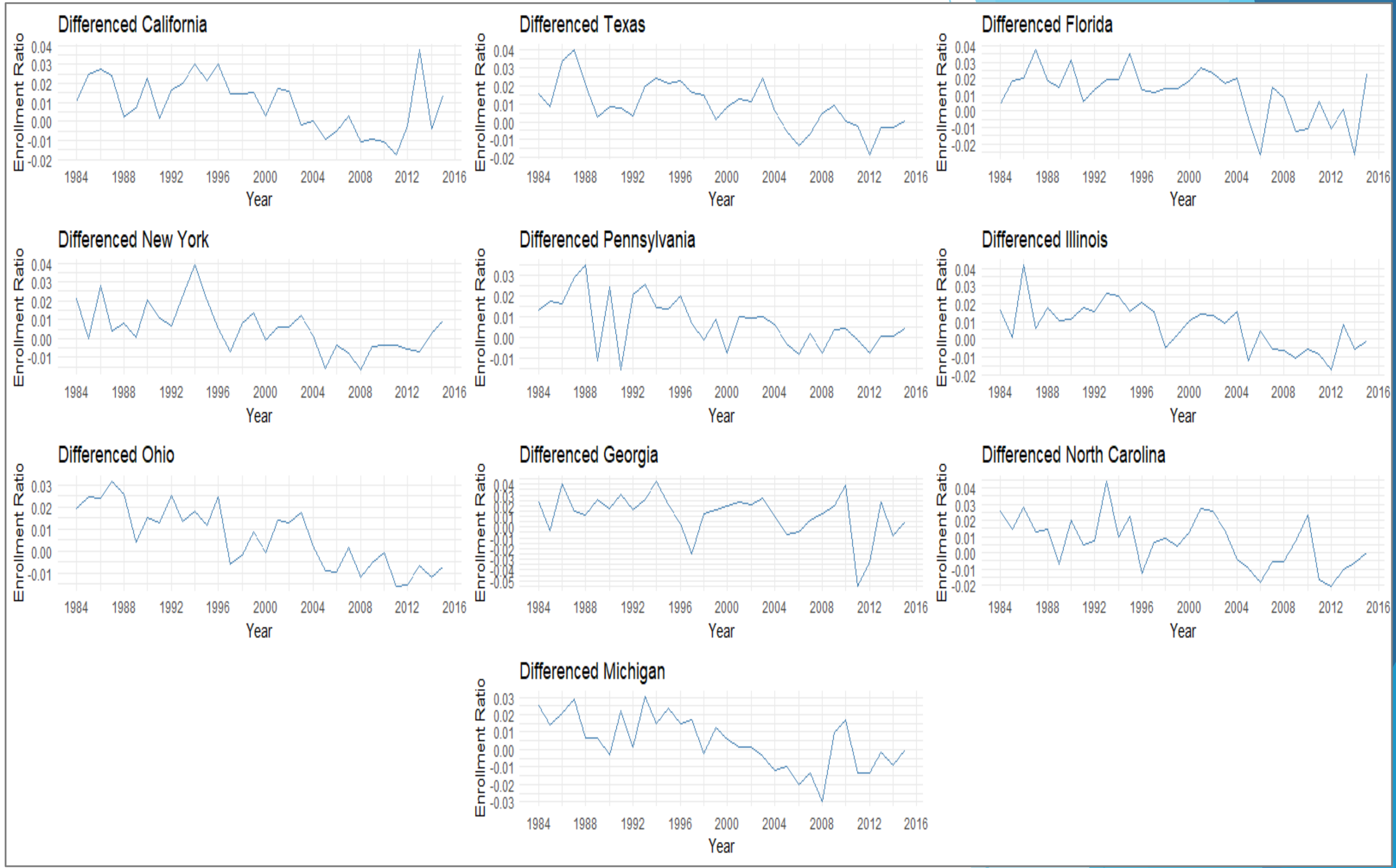


Augmented Dickey – Fuller Table

	OVERALL VALUE OF THE TEST STATISTIC			5 PCT CRITICAL VALUE			SIGNIFICANCE LEVEL	REJECT/FAIL TO REJECT
	Tau	**Phi1**	**Phi2**	**Tau**	**Phi1**	**Phi2**	**P - Value**	**Null**
CAL (NO TREND OR DRIFT)	1.6339	N/A	N/A	-1.95	N/A	N/A	0.0003612	Reject
CAL (WITH DRIFT)	-2.5808	5.4213	N/A	-2.93	4.86	N/A	0.002575	Reject
CAL (WITH TREND)	-1.3029	3.4873	3.2143	-3.50	5.13	6.73	0.008465	Reject
TEX (NO TREND OR DRIFT)	0.9671	N/A	N/A	-1.95	N/A	N/A	0.000001416	Reject
TEX (WITH DRIFT)	-2.5757	4.1546	N/A	-2.93	4.86	N/A	0.0000179	Reject
TEX (WITH TREND)	-1.0167	2.679	3.2104	-3.50	5.13	6.73	0.0000845	Reject
FLO (NO TREND OR DRIFT)	1.9751	N/A	N/A	-1.95	N/A	N/A	0.0007329	Reject
FLO (WITH DRIFT)	-2.9445	7.7268	N/A	-2.93	4.86	N/A	0.004278	Reject
FLO (WITH TREND)	0.0534	5.6747	5.0759	-3.50	5.13	6.73	0.007176	Reject
NY (NO TREND OR DRIFT)	1.0558	N/A	N/A	-1.95	N/A	N/A	0.005209	Reject
NY (WITH DRIFT)	-2.028	2.8054	N/A	-2.93	4.86	N/A	0.0063	Reject
NY (WITH TREND)	-0.8408	2.033	2.3118	-3.50	5.13	6.73	0.01461	Reject
PENN (NO TREND OR DRIFT)	2.1085	N/A	N/A	-1.95	N/A	N/A	0.02229	Reject
PENN (WITH DRIFT)	-3.9875	12.153	N/A	-2.93	4.86	N/A	0.001549	Reject
PENN (WITH TREND)	-1.3886	8.1307	8.0665	-3.50	5.13	6.73	0.004202	Reject
ILL (NO TREND OR DRIFT)	1.4062	N/A	N/A	-1.95	N/A	N/A	0.002533	Reject
ILL (WITH DRIFT)	-2.8864	5.8262	N/A	-2.93	4.86	N/A	0.002718	Reject
ILL (WITH TREND)	0.2824	6.248	7.4563	-3.50	5.13	6.73	0.000889	Reject
OH (NO TREND OR DRIFT)	0.4805	N/A	N/A	-1.95	N/A	N/A	0.00001131	Reject
OH (WITH DRIFT)	-2.8793	4.3976	N/A	-2.93	4.86	N/A	0.000004567	Reject
OH (WITH TREND)	-0.5345	6.218	9.0138	-3.50	5.13	6.73	0.0000008396	Reject
GEO (NO TREND OR DRIFT)	1.8745	N/A	N/A	-1.95	N/A	N/A	0.02464	Reject
GEO (WITH DRIFT)	-2.0681	4.6521	N/A	-2.93	4.86	N/A	0.09245	Fail to Reject
GEO (WITH TREND)	-0.8528	3.0192	2.0992	-3.50	5.13	6.73	0.1918	Fail to Reject
NC (NO TREND OR DRIFT)	1.1464	N/A	N/A	-1.95	N/A	N/A	0.01647	Reject
NC (WITH DRIFT)	-2.3045	3.6038	N/A	-2.93	4.86	N/A	0.01232	Reject
NC (WITH TREND)	-0.6531	2.423	2.7113	-3.50	5.13	6.73	0.03084	Reject
MIC (NO TREND OR DRIFT)	0.5588	N/A	N/A	-1.95	N/A	N/A	0.009935	Reject
MIC (WITH DRIFT)	-2.1207	2.4969	N/A	-2.93	4.86	N/A	0.003394	Reject
MIC (WITH TREND)	-0.9705	3.0508	4.3043	-3.50	5.13	6.73	0.002124	Reject

First Differenced Plots

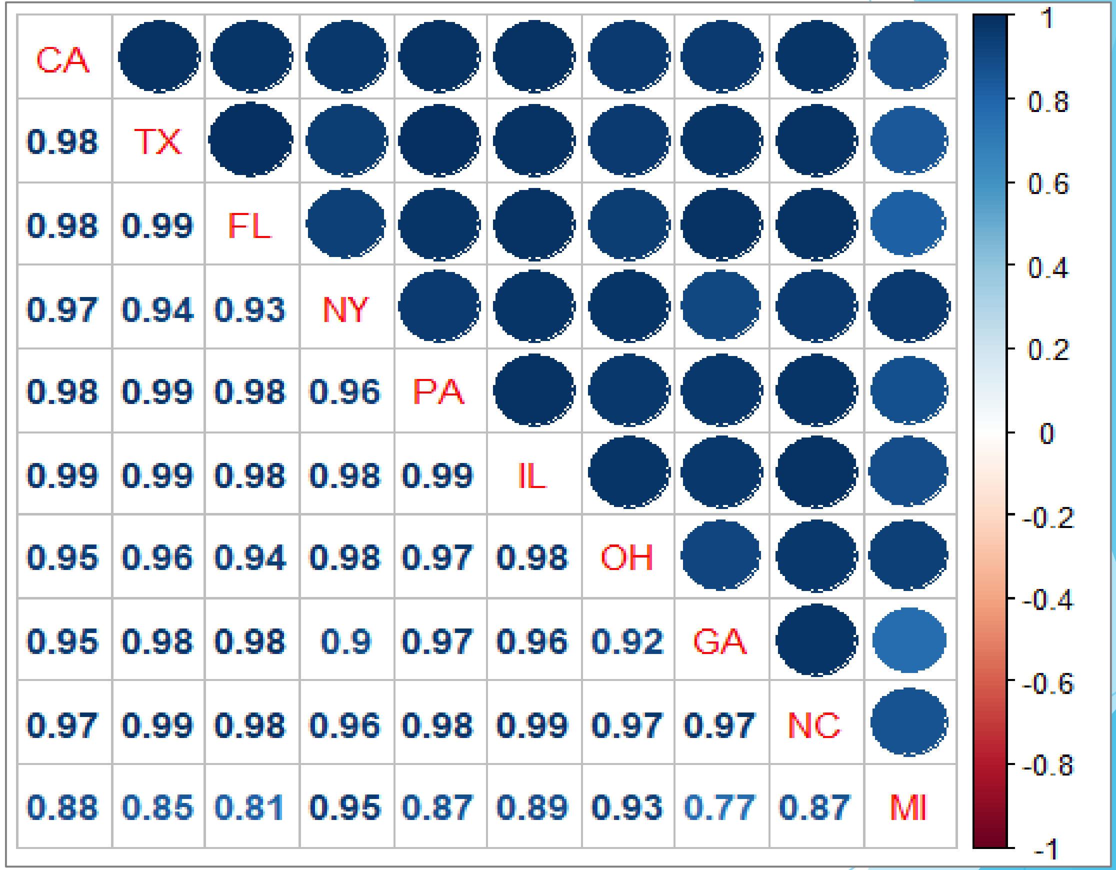
- Since Georgia has been shown to be nonstationary, the rest of the data is first differenced so that the general comparisons can be made
- Differencing thus gives the following graphs
- After differencing, it appears that these follow some very different paths, though there could be errors shown here that are not taken into account by visual analysis
- Keep in mind that each of these are scaled slightly differently, due to the fact that they each move a bit differently
- The differences in these plots will be better shown by a multivariate time series plot, which will be shown later



```
g1cald <- ggplot(data = data6, mapping = aes(x = Year, y = DiffCalifornia, color = Value)) +
  ggtitle("Differenced California") +
  xlab("Year") +
  ylab("Enrollment Ratio") +
  geom_line(stat="identity", color = "steelblue", show.legend = TRUE) +
  theme_minimal() +
  scale_y_continuous(breaks = round(seq(min(-0.03), max(0.04), by = 0.01),2)) +
  scale_x_continuous(breaks = round(seq(min(1984), max(2016), by = 4),1)) +
  expand_limits(x=1984, y=0)
g1cald
```

Correlation Plot

- Most of the correlations stay between a value of 0.97 and 0.97
- The two most different here are Michigan and Georgia, with a low of 0.77, where Michigan appears to follow a very different process than the other nine
- Besides Michigan and Georgia, the only other state that could be following strange patterns seems to be New York, since it and Florida have a correlation of only 0.93

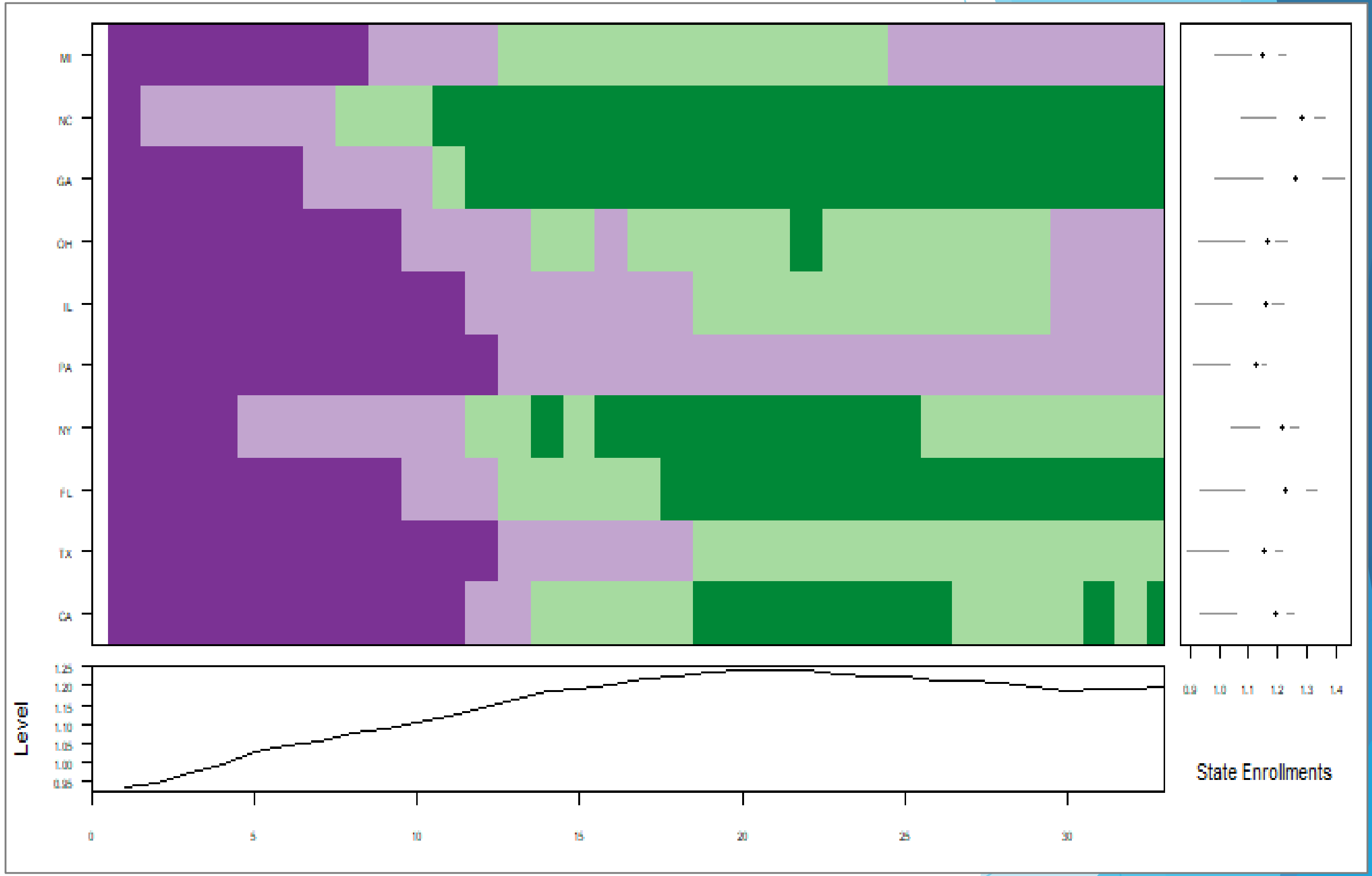


CA	●	●	●	●	●	●	●	●	●
0.98	TX	●	●	●	●	●	●	●	●
0.98	0.99	FL	●	●	●	●	●	●	●
0.97	0.94	0.93	NY	●	●	●	●	●	●
0.98	0.99	0.98	0.96	PA	●	●	●	●	●
0.99	0.99	0.98	0.98	0.99	IL	●	●	●	●
0.95	0.96	0.94	0.98	0.97	0.98	OH	●	●	●
0.95	0.98	0.98	0.9	0.97	0.96	0.92	GA	●	●
0.97	0.99	0.98	0.96	0.98	0.99	0.97	0.97	NC	●
0.88	0.85	0.81	0.95	0.87	0.89	0.93	0.77	0.87	MI

```
# Correlation Plot
library("corrplot")
corrplot.mixed(cor(data5))
```

Multivariate Time Series Plot
(Developed by Roger Peng 2008)

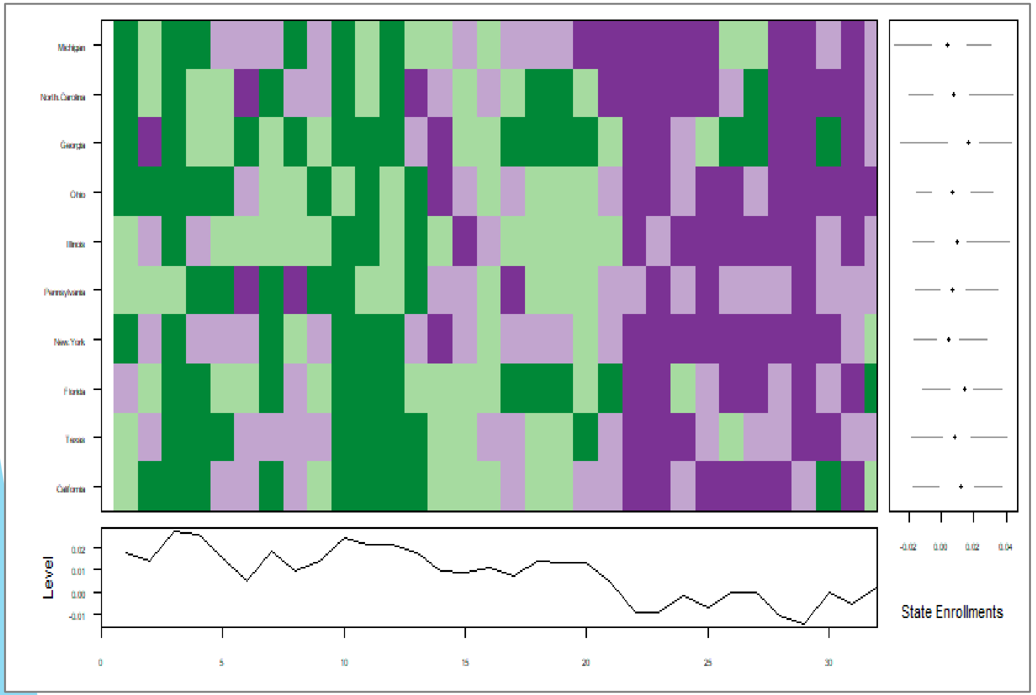
- This plot is read and understood similar to a heat map
- The variables are plotted on the y – axis and the time periods on the x – axis
- The median for each point in time is plotted below
- The median and boxplot information are given on the side
  - The placement of the dot shows the median
  - The length of the line on either side of the dot shows the length of the data
- The colors are divided into the quantiles of the averaged dataset
- Here,
  - Median Range: 1.1237 (Pennsylvania) to 1.281 (North Carolina)
  - Data is set into quantiles
    - Dark Green: High
    - Light Green: Medium High
    - Light Purple: Medium Low
    - Dark Purple: Low
  - Basically, this visualization shows the distribution of the data
- Here, there is a lot of clustering, though that could be due to the trend
- It seems that the median and whiskers of the states do not match up at all



```
library("mvtsplot")
datmat1 <- data.matrix(data5)
mvtsplot(datmat1, group = NULL, xtime = NULL, norm = "global", levels = 4,
        smooth.df = NULL, margin = TRUE, sort = NULL, main = "State Enrollments",
        palette = "PRGn", rowstat = "median", xlim, bottom.ylim = NULL,
        right.xlim = NULL, gcol = 1)
```

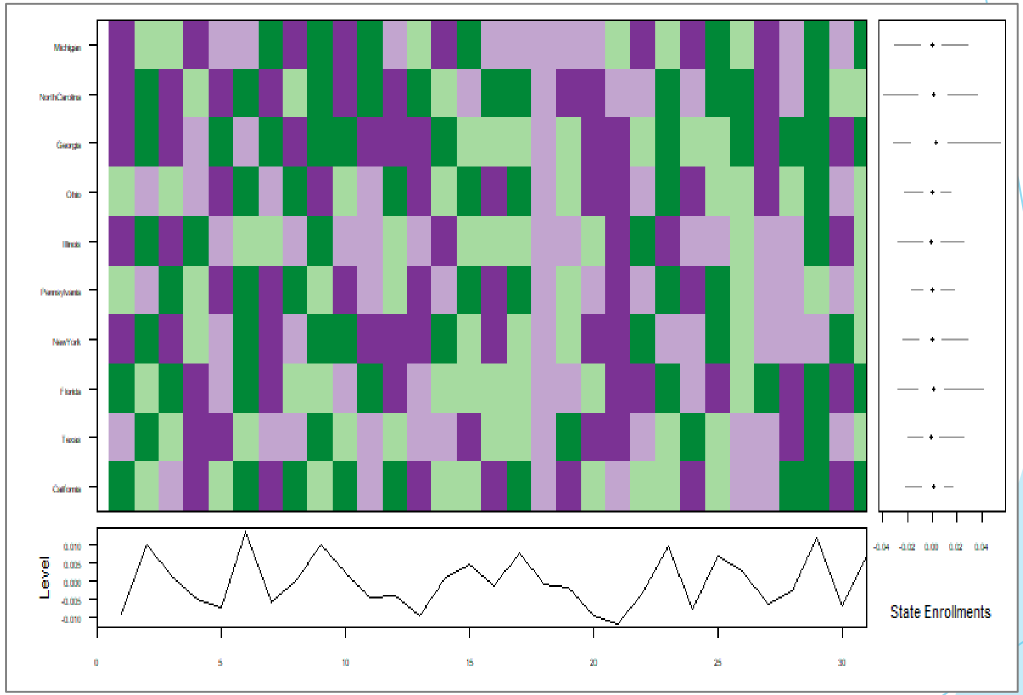
More Accurate Multivariate Time Series Plots

Differenced MVTS Plot



- Dark Green: High
- Light Green: Medium High
- Light Purple: Medium Low
- Dark Purple: Low
- Median Range: 0.004449 (Michigan) to 0.01631 (Georgia)

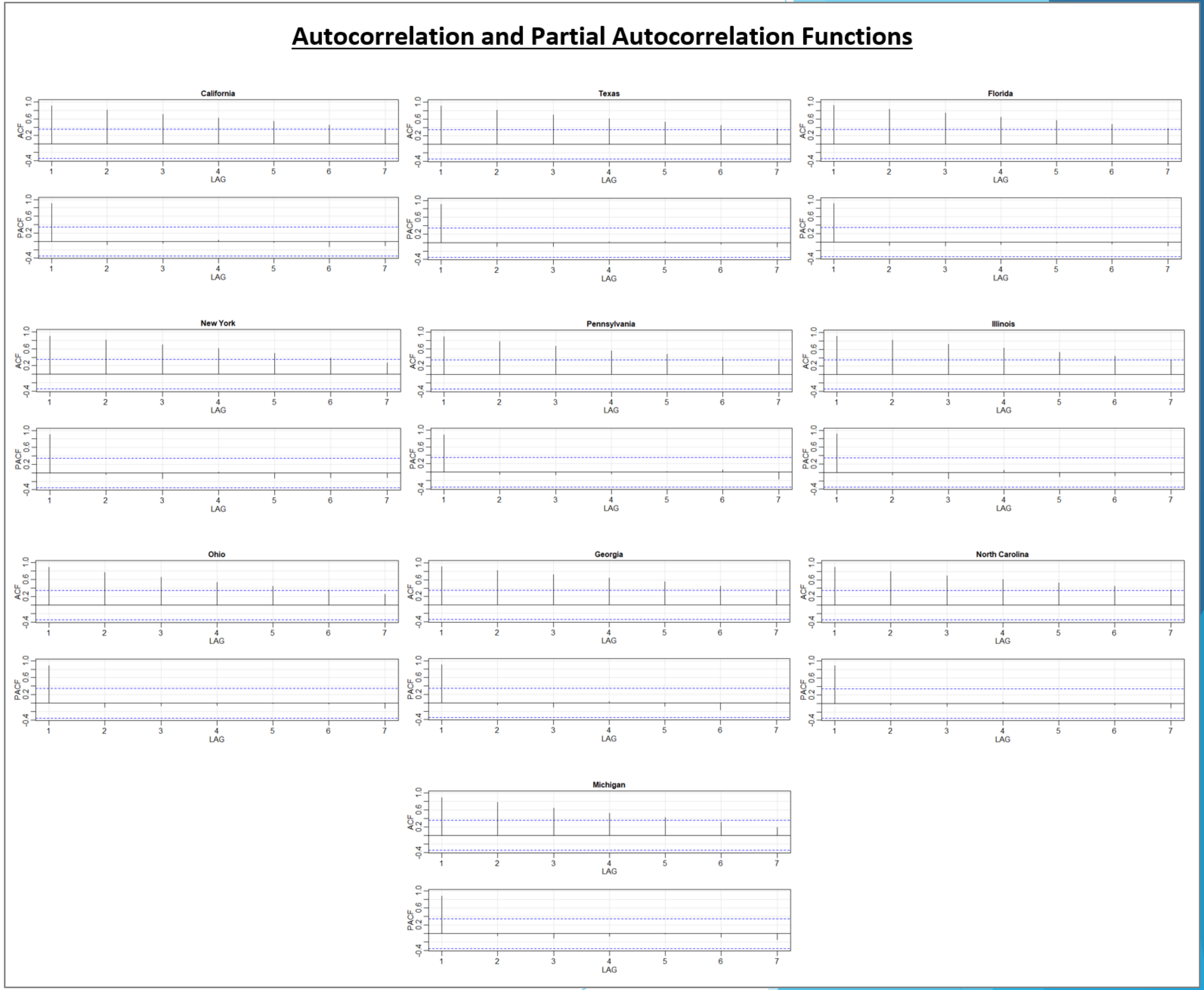
Twice Differenced MVTs Plot



- Dark Green: High
- Light Green: Medium High
- Light Purple: Medium Low
- Dark Purple: Low
- Median Range: -0.0010250 (Illinois) to 0.0033709 (Georgia)

Autocorrelation and Partial Autocorrelation Functions

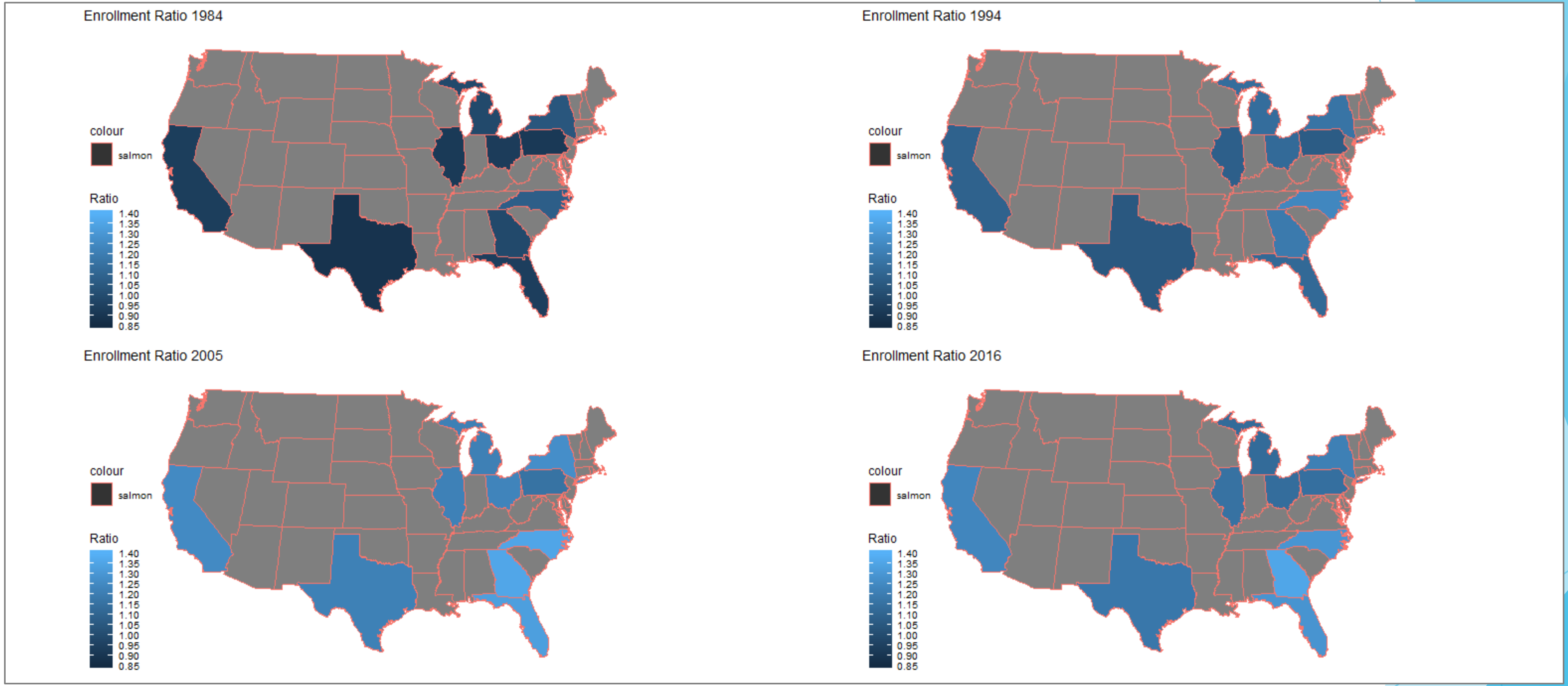
- The dynamics of each state look fairly similar, with some persistence in the autocorrelation functions
- This persistence may suggest the presence of a moving average process present within the data, since there are a single spike for each of the partial autocorrelation functions
  - Specifically, they all look like MA(1) processes
- It is interesting to note that all of the autocorrelation and partial autocorrelation functions are so similar, suggesting that they may obey the same dynamics
- Given more data, it could be theoretically possible that the autocorrelation persistence shows trend, but, since the Dickey – Fuller tests show that only Georgia needs to be first – differenced, it cannot be assumed here that this is the case



Autocorrelation and Partial Autocorrelation Functions

```
library("astsa")
acf2(California, main = "California")
```

Geographically Speaking…


Enrollment Ratio 1984

Enrollment Ratio 1994

Enrollment Ratio 2005

Enrollment Ratio 2016

- These four plots show the enrollment ratios within the ten states over time
  - None of the data is differenced for these four maps
- Many of the states become lighter until 2005
- In 2016, some of the states becomes darker
- It does appear that there may be some geographic differences within the data that influence it over time
  - North Carolina, Georgia, and Florida seem to move mostly together, and the northeast moves together as well

Geographically Speaking…

```
library("maps")
library("dplyr")
library("ggplot2")
library("gridExtra")

dfstate <- map_data("state")
ggplot(dfstate) + geom_polygon(aes(long, lat, group = group), color = "white")
aggdat1 <- left_join(dfstate, mapdata1, by = c("region" = "state"))

map2016 <- ggplot(data = aggdat1) +
  ggtitle("Enrollment Ratio 2016") +
  labs(fill = "Enrollment Ratio") +
  geom_polygon(aes(long, lat, group = group, fill = enrollrat, color = "salmon")) +
  coord_map("bonne", parameters = 10) +
  ggthemes::theme_map() +
  scale_fill_continuous(limits = c(0.85,1.40), breaks = c(0.85, 0.90, 0.95, 1.00, 1.05, 1.10, 1.15, 1.20, 1.25, 1.30, 1.35, 1.40),
  guide = guide_colourbar(title = "Ratio", draw.ulim = FALSE, draw.llim = FALSE))
Map2016

grid.arrange(map1984, map1994, map2005, map2016)
```

Michigan and Georgia

- It seems that no matter which method is used or which way the data is plotted, Michigan and Georgia keep deviating slightly from the normal ranges
- Though this analysis is to determine if there is a possibility that the aggregate dataset does not fully capture the underlying processes of the state datasets and to examine the general patterns within the state datasets, some possibilities as to why Michigan and Georgia are different are suggested and recommended for further analysis
    - One reason that Michigan and Georgia may deviate from the regular pattern could be regional or cultural factors
        - Seen before in the previous graph
        - Georgia is in the southeast right above Florida, while Michigan is in the northeast surrounded by the Great Lakes
            - This could, in theory, attract different kinds of students, along with political factors stemming from these
    - Previous research can be used to answer why these two are so vastly different
        - As stated before, in 1996 Jerry A. Jacobs found that larger schools (engineering and technology, specifically) have not accepted as many females, while liberal arts schools, among others, have accepted females in large numbers
        - Both Michigan and Georgia are known for their tech schools, though they do both have liberal arts schools
        - Many of the other states, while they do have tech schools, they are not as prevalent
        - Furthermore, Georgia released a plan in August 2011 called “Complete College Georgia Initiative” (Governor’s Office of Student Achievement 2011)

Conclusion

- While the overall dataset appears to reflect the same patterns as most of the states, it does not appear to capture the differing patterns that are occurring in states such as Michigan and Georgia.
- Some limitations to this study exist
  - Though this study seeks to understand if the enrollment rates differ throughout the years, no empirical way was found to directly test this phenomenon
  - Though an SUR model could be set up and linear dynamic regressions performed, utilizing Theil's F – test or other methods would only test the possibility of differences within the coefficients, and not the processes themselves
    - For this reason, only the univariate processes were analyzed
    - It is also not possible to test univariate processes through the SUR model, and there does not seem to be many ways to test this besides direct comparison
    - It is highly probable that other multilevel models could do a better job at showing the differences within the data from the aggregated dataset
  - Only 33 data points have been provided for analysis, and the year 2002 was never reported, thus giving some of the data higher standard errors that make arima and other models difficult to fit
    - In later years, it may be more possible to see if the data follows the same sorts of patterns
- For future studies, it may be better to gather more data and find a way to run an F test on the univariate model.

Bibliography

- Agger, Charlotte, Judith Meece, and Soo-yong Byun. 2018. “The Influences of Family and Place on Rural Adolescents’ Educational Aspirations and Post-secondary Enrollment.” *Journal of Youth and Adolescence* 47 (December 2018): 2554 – 2568.
- Conger, Dylan and Mark C. Long. 2010. “Why Are Men Falling Behind? Gender Gaps in College Performance and Persistence.” *The ANNALS of the American Academy of Political and Social Science* 627, no. 1 (January 4): 184 – 214.
- National Center for Education Statistics. (1990-2017). “Total Fall Enrollment in Degree-Granting Postsecondary Institutions, by Attendance Status, Sex, and State or Jurisdiction”. *Digest of Education Statistics.* Accessed at: https://nces.ed.gov/pubsearch/getpubcats.asp?sid=091#061
- National Center for Education Statistics. (1987-1989). “Total Fall Enrollment in Degree-Granting Postsecondary Institutions, by Attendance Status, Sex, and State or Jurisdiction”. *Digest of Education Statistics.* Accessed at: https://catalog.hathitrust.org/api/volumes/oclc/3133477.html
- Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko. 2006. “The Homecoming of American College Women: The Reversal of the College Gender Gap.” *Journal of Economic Perspectives* 20 no. 4 (Fall): 133-156.
- Jacobs, Jerry A. 1996. “Gender Inequality and Higher Education.” *Annual Review of Sociology* 22: 153-185.
- Lopez, Mark Hugo and Ana Gonzalez-Barrera. 2014. “Women’s College Enrollment Gains Leave Men Behind.” *Pew Research Center*, March 6. http://www.pewresearch.org/fact-tank/2014/03/06/womens-college-enrollment-gains-leave-men-behind/ (February 19, 2019).
- Peng, Roger. 2008. “A Method for Visualizing Multivariate Time Series Data.” *Journal of Statistical Software* 25 (March 31): 1 – 17.
- The Governor’s Office of Student Achievement. 2011. “Complete College Georgia: An Overview.” https://gosa.georgia.gov/complete-college-georgia-overview (April 30, 2019).