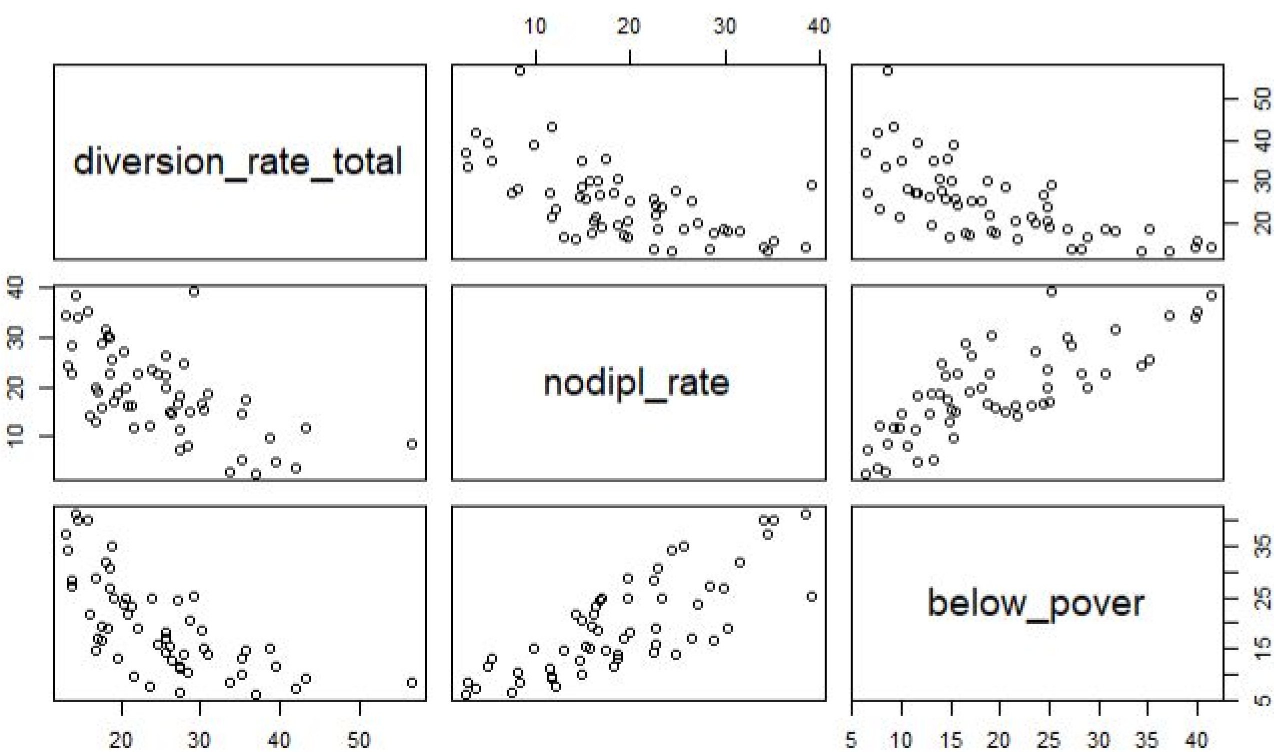
Project 2: Identifying Clusters of NYC Municipal Sub-areas by Solid Waste Diversion Rates with Group Based Trajectory Models

Introduction and Data

New York City's solid waste recycling program was initiated in 1988 in a couple of neighborhoods and overtime became expanded citywide. Although today the program is uniform in implementation in every neighborhood, solid waste diversion rates vary significantly throughout the city's municipal subdivisions, ranging from 56 to 13% in 2017.

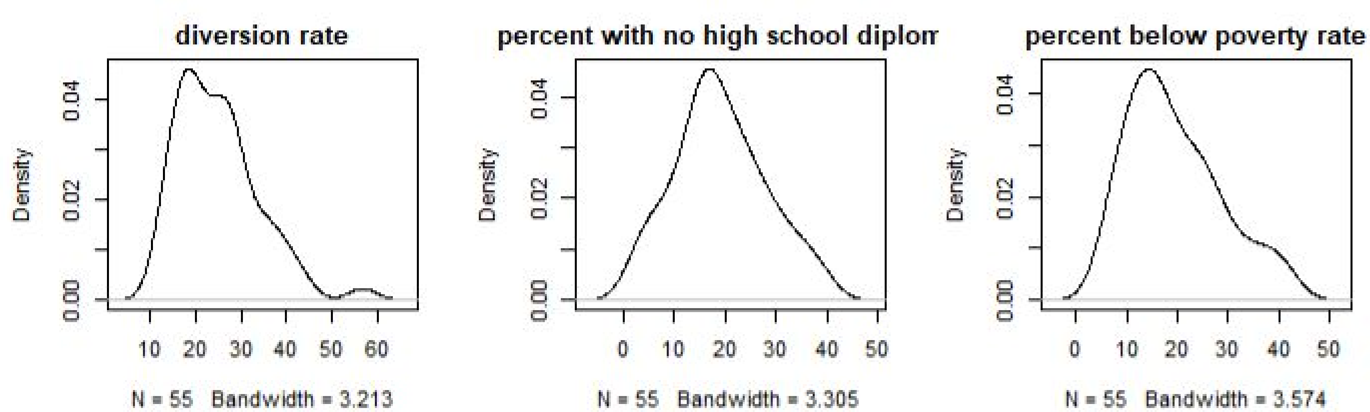
This paper explores the possible group trajectories of NYC sub-areas based on their diversion rates from 2010 - 2017. Recycling rates for a sample of 55 Public Use Microdata Areas (PUMA) similar in geography to community districts were drawn from the NYC Open Data resource (https://data.cityofnewyork.us/City-Government/DSNY-Monthly-Tonnage-Data/ebb7-mvp5), to which were joined socio-economic characteristics (percent of population with no high school diploma and percent of population below poverty level) collected as part of the American Community Survey and available on the American Fact Finder website https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

The paper also seeks to present a profile of identified clusters by using variables that are known to correlate strongly with diversion rates: level of education and economic status, as presented on the graph below.



Analysis

I explored the need to transform or rescale the measurements by examining univariate densities below.



The densities are sufficiently symmetric, so no transformations were undertaken.
Although the measurements are on somewhat different scales, I decided not to normalize the data as the clusters with normalized data were quite similar to the ones done with raw data, while interpretation with raw data is a little more straightforward.

Clustering Methods Used

Two mixed model clustering methods for longitudinal outcomes were implemented: Nagin clustering and the mixed models for Gaussian longitudinal outcomes using hlme function.

I successively examined linear, quadratic and cubic models for two and three clusters. I did not increase the number of clusters beyond three as significance levels of coefficients began to drop. I also constructed models with additional time-dependent predictors: percent of population below poverty rate and percent of population with no high school diploma.

BIC was used to guide the choice between the models.

BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. [1]

[1] https://methodology.psu.edu/AIC-vs-BIC

Model 1: linear mixed model with two clusters

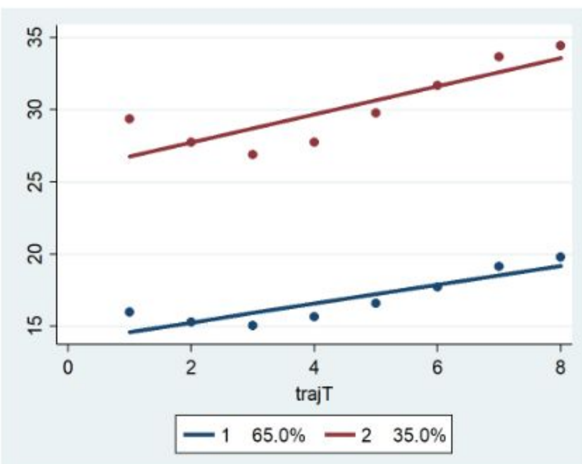
```
%%stata
use "final_1.dta", clear
traj , model(cnorm) var(Y2010 Y2011 Y2012 Y2013 Y2014 Y2015 Y2016 Y2017) indep(X1-X8) order(1 1) min(-100) max(100)
trajplot
```

```
### latent class linear mixed model with 2 classes

mod_1<-hlme(diversion_rate_total~time, mixture=~time,random=~time, subject='puma',
        ng=2,data=data_long)

summary(mod_1)
```

Maximum Likelihood Estimates
Model: Censored Normal (cnorm)

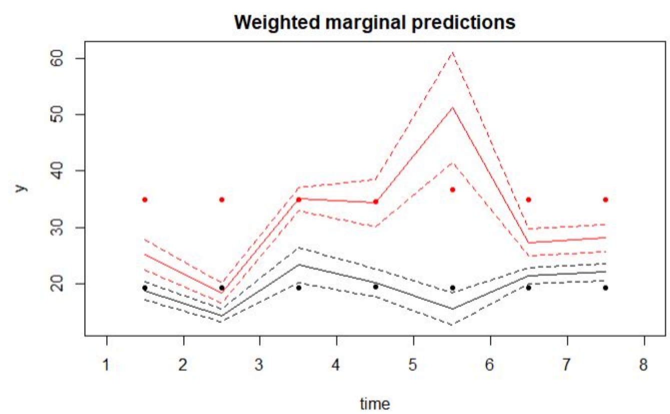
Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	13.97424	0.65976	21.181	0.00
	Linear	0.65268	0.12840	5.083	0.00
2	Intercept	25.80238	0.90101	28.637	0.00
	Linear	0.97067	0.17578	5.522	0.00
	Sigma	4.94436	0.16843	29.355	0.00
Group membership					
1	(%)	65.04514	6.57288	9.896	0.00
2	(%)	34.95486	6.57288	5.318	0.00

BIC= -1380.30 (N=440) BIC= -1374.06 (N=55) AIC= -1368.04 ll= -1362.04




```
Goodness-of-fit statistics:
     maximum log-likelihood: -984.94
     AIC: 1987.88
     BIC: 2005.95


Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                     coef      Se    Wald p-value
intercept class1  1.78347 0.51065  3.493 0.00048

Fixed effects in the longitudinal model:

                     coef      Se    Wald p-value
intercept class1 15.93572 0.85548 18.628 0.00000
intercept class2 31.03937 2.90355 10.690 0.00000
time class1       0.74537 0.07319 10.184 0.00000
time class2       0.87369 0.21111  4.139 0.00003
```

```{r}
plot(mod_1,which='fit',var.time='time',legend=NULL)
```



The coefficients produced by both traj and hlmn are similar. The level and slope are significant in both clusters in both models. BIC is somewhat larger in the hlme model.

Model 2: linear mixed model with three clusters

%%stata
use "final_1_wide.dta", clear
traj , model(cnorm) var(Y2010 Y2011 Y2012 Y2013 Y2014 Y2015 Y2016 Y2017) indep(X1-X8) order(1 1 1) min(-100) max(100)
trajplot

```
```{r}
mod_2<-hlme(diversion_rate_total~time, mixture=~time,random=~time, subject='puma',
         ng=3,data=data_long)

summary(mod_2)
```

```
                         Maximum Likelihood Estimates
                         Model: Censored Normal (cnorm)

                                          Standard       T for H0:
 Group   Parameter          Estimate        Error     Parameter=0   Prob > |

 1       Intercept          11.83916      0.53729          22.035      0.00
         Linear              0.64921      0.10554           6.152      0.00

 2       Intercept          20.18462      0.58137          34.719      0.00
         Linear              0.79227      0.11479           6.902      0.00

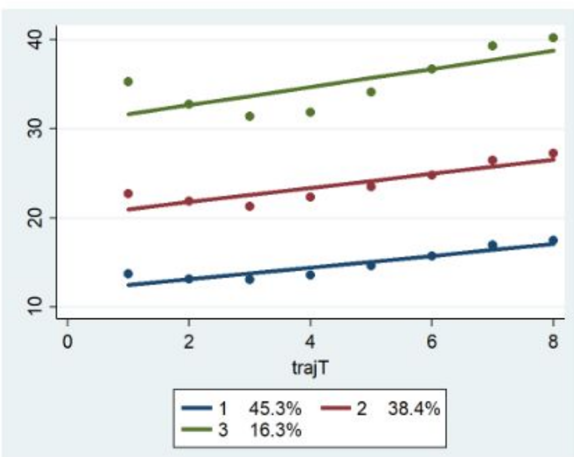
 3       Intercept          30.60750      0.89625          34.151      0.00
         Linear              1.01471      0.17586           5.770      0.00

         Sigma               3.38021      0.11531          29.314      0.00

  Group membership

 1       (%)                45.26469      6.80218           6.654      0.00
 2       (%)                38.40898      6.65278           5.773      0.00
 3       (%)                16.32633      5.03762           3.241      0.00

 BIC= -1243.34 (N=440)  BIC= -1233.99 (N=55)  AIC= -1224.95  ll=  -1215.95
```




```
Goodness-of-fit statistics:
     maximum log-likelihood: -979.85
     AIC: 1983.71
     BIC: 2007.8

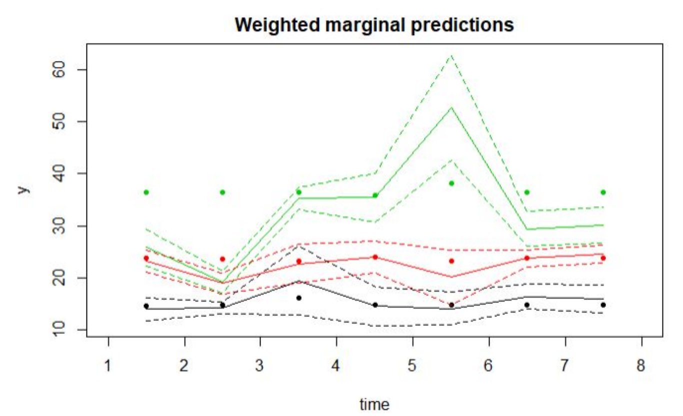

Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                    coef      Se    Wald p-value
intercept class1  1.20424 0.44997  2.676 0.00744
intercept class2  1.27551 0.44371  2.875 0.00404

Fixed effects in the longitudinal model:

                    coef      Se    Wald p-value
intercept class1 11.45848 0.75234 15.231 0.00000
intercept class2 20.23608 0.73626 27.485 0.00000
intercept class3 32.66536 1.05296 31.022 0.00000
time class1       0.72982 0.10593  6.890 0.00000
time class2       0.76966 0.10291  7.479 0.00000
time class3       0.85636 0.18189  4.708 0.00000
```

```
```{r}
plot(mod_2,which='fit',var.time='time',legend=NULL)
```
```




The coefficients produced by both traj and hlmn are similar. The level and slope are significant in the three clusters in both models.
The clusters in the three cluster model are sufficiently well defined. However, BIC has decreased in the 3-cluster traj model while it has increased slightly in the hlme model. I decided not to run models with greater number of clusters as clusters become less clearly defined graphically and significance levels of the coefficients decrease.
```

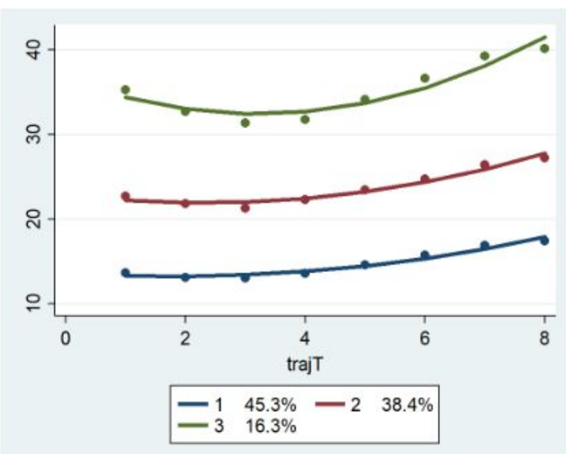
Models with higher order terms

Model 3: three clusters with quadratic mean

```
%%stata
use "final_1_wide.dta", clear
traj , model(cnorm) var(Y2010 Y2011 Y2012 Y2013 Y2014 Y2015 Y2016  Y2017) indep(X1-X8) order(2 2 2) min(-100) max(100)
trajplot
```

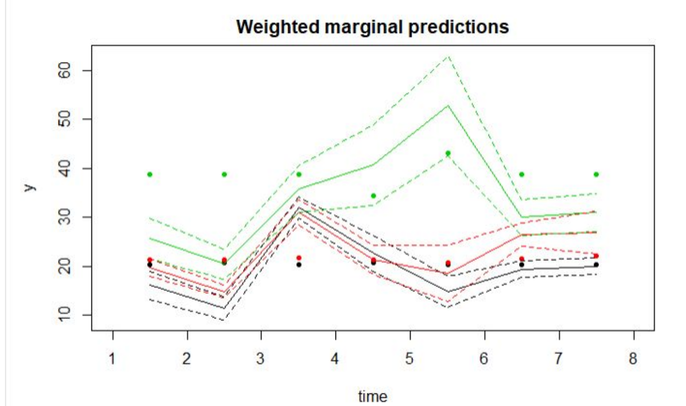
```
mod_4 <- hlme(diversion_rate_total~time+I(time^2),
              mixture=~time+I(time^2),
              classmb=~1, random=~1,subject='puma',ng=3,
              data=data_long)
```

```
                         Maximum Likelihood Estimates
                         Model: Censored Normal (cnorm)

                                            Standard       T for H0:
 Group   Parameter        Estimate        Error     Parameter=0   Prob > |

 1       Intercept        13.68035      0.91875        14.890        0.00
         Linear           -0.45359      0.46707        -0.971        0.33
         Quadratic         0.12252      0.05066         2.419        0.01

 2       Intercept        22.93646      0.99715        23.002        0.00
         Linear           -0.85766      0.50746        -1.690        0.09
         Quadratic         0.18333      0.05504         3.331        0.00

 3       Intercept        36.53522      1.53072        23.868        0.00
         Linear           -2.54467      0.77792        -3.271        0.00
         Quadratic         0.39552      0.08436         4.688        0.00

         Sigma             3.23589      0.11072        29.225        0.00

  Group membership

 1       (%)              45.29626      6.82116         6.641        0.00
 2       (%)              38.36577      6.66782         5.754        0.00
 3       (%)              16.33797      5.05400         3.233        0.00

 BIC= -1233.37 (N=440)  BIC= -1220.89 (N=55)  AIC= -1208.84  ll=  -1196.84
```




```
Goodness-of-fit statistics:
     maximum log-likelihood: -874.08
     AIC: 1774.15
     BIC: 1800.25


Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                    coef      Se    Wald p-value
intercept class1  2.64816 0.73509  3.603 0.00032
intercept class2  2.51319 0.73953  3.398 0.00068

Fixed effects in the longitudinal model:

                    coef      Se    Wald p-value
intercept class1 21.10177 1.44709 14.582 0.00000
intercept class2 19.43162 1.54693 12.561 0.00000
intercept class3 37.97642 5.12813  7.406 0.00000
time class1      -1.09812 0.15752 -6.971 0.00000
time class2      -0.59935 0.17054 -3.514 0.00044
time class3      -3.19280 0.56838 -5.617 0.00000
I(time^2) class1  0.16803 0.01707  9.844 0.00000
I(time^2) class2  0.18330 0.01834  9.997 0.00000
I(time^2) class3  0.59546 0.06179  9.637 0.00000
```

```{r}
plot(mod_4, which='fit',var.time='time',legend=NULL)
```



BIC for both models became smaller, all coefficients in both models are significant.

Linear term in the first group of the traj model is less significant suggesting that one polynomial is probably insufficient, while the traj model shows that group 1's linear model is insignificant and group 2's linear model is less significant, indicating that that the quadratic term is probably more appropriate.
While the coefficients produced by the hlme function are all significant, the graph shows that there is a significant and continuous overlap of two clusters.

Model 4: three clusters with cubic mean

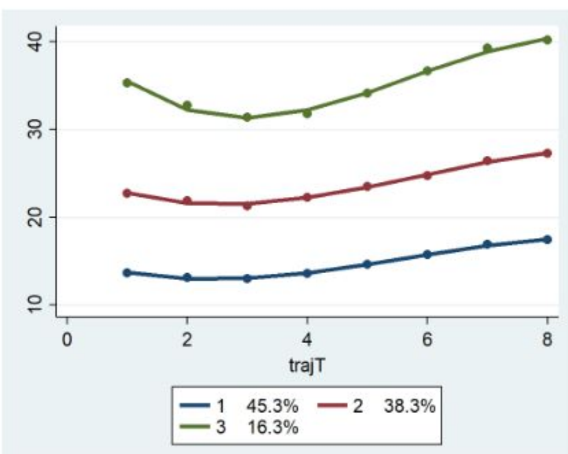
```
%%stata
use "final_1_wide.dta", clear
traj , model(cnorm) var(Y2010 Y2011 Y2012 Y2013 Y2014 Y2015 Y2016 Y2017) indep(X1-X8) order(3 3 3) min(-100) max(100)
trajplot
```

```r
mod_5 <- hlme(diversion_rate_total~time+I(time^3),
              mixture=~time+I(time^3),
              classmb=~1, random=~1,subject='puma',ng=3,
              data=data_long)
```

Model: Censored Normal (cnorm)

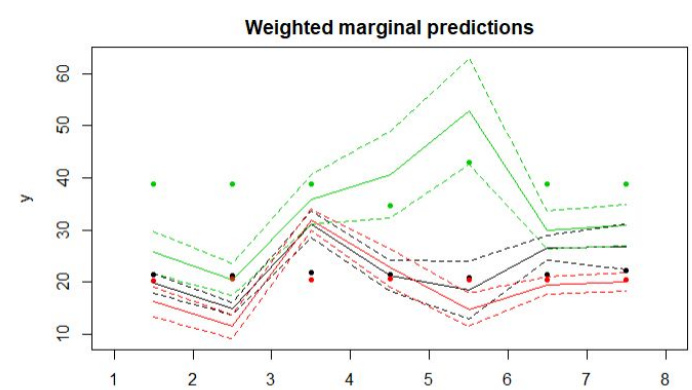
Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	15.50494	1.60423	9.665	0.00
	Linear	-2.35036	1.45122	-1.620	0.10
	Quadratic	0.61970	0.36404	1.702	0.08
	Cubic	-0.03683	0.02671	-1.379	0.16
2	Intercept	25.29649	1.74604	14.488	0.00
	Linear	-3.31237	1.57792	-2.099	0.03
	Quadratic	0.82681	0.39584	2.089	0.03
	Cubic	-0.04767	0.02904	-1.641	0.10
3	Intercept	41.83101	2.67235	15.653	0.00
	Linear	-8.05730	2.41609	-3.335	0.00
	Quadratic	1.84068	0.60603	3.037	0.00
	Cubic	-0.10705	0.04446	-2.408	0.01
	Sigma	3.19685	0.10975	29.127	0.00
Group membership					
1	(%)	45.30956	6.84324	6.621	0.00
2	(%)	38.34645	6.68780	5.734	0.00
3	(%)	16.34399	5.07129	3.223	0.00

BIC= -1237.19 (N=440) BIC= -1221.60 (N=55) AIC= -1206.54 ll= -1191.54




```
Goodness-of-fit statistics:
     maximum log-likelihood: -893.89
     AIC: 1813.78
     BIC: 1839.87

Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                    coef      Se    Wald p-value
intercept class1  2.50609 0.74136  3.380 0.00072
intercept class2  2.65469 0.73607  3.607 0.00031

Fixed effects in the longitudinal model:

                    coef      Se    Wald p-value
intercept class1 18.49356 1.56618 11.808 0.00000
intercept class2 20.34806 1.45610 13.974 0.00000
intercept class3 35.53135 5.09955  6.968 0.00000
time class1       0.21019 0.11069  1.899 0.05758
time class2      -0.38576 0.10186 -3.787 0.00015
time class3      -0.80072 0.35810 -2.236 0.02535
I(time^3) class1  0.01200 0.00143  8.368 0.00000
I(time^3) class2  0.01148 0.00132  8.682 0.00000
I(time^3) class3  0.04239 0.00478  8.867 0.00000


Variance-covariance matrix of the random-effects:
          intercept
intercept  50.02386

                            coef      Se
Residual standard error:  1.18904 0.04409
```

```{r}
plot(mod_5, which='fit',var.time='time',legend=NULL)
```



The cubic terms produced by the traj model are not significant. While the hlme method's groups appear to have significant coefficients, the graph shows significant overlap between the group. T

Models with time-dependent predictors

Model 5: three clusters with quadratic mean and percent of population with no high school diploma

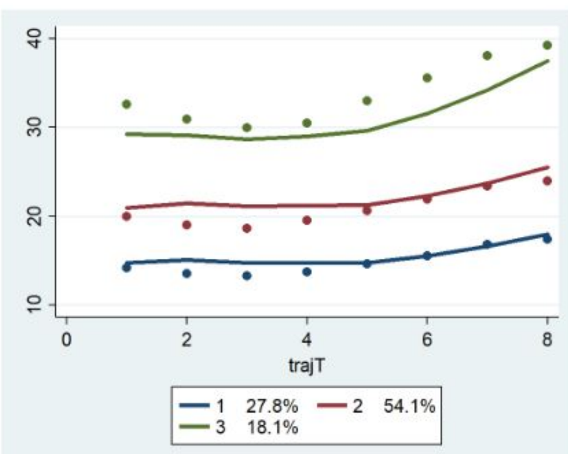
```
%%stata
use "final_1_wide.dta", clear
traj , model(cnorm) var(Y2010 Y2011 Y2012 Y2013 Y2014 Y2015 Y2016 Y2017) indep(X1-X8) order(2 2 2) tcov(nd0 nd1 nd2 nd3 nd4 nd5 nd6 nd7) min(-100) max(100)
trajplot
```

~~~
```{r}
mod_6 <- hlme(diversion_rate_total~time+I(time^3),
                  mixture=~time+I(time^3),
                  classmb=~nodipl_rate, random=~1,subject='puma',ng=3,
                  data=data_long)
~~~

Maximum Likelihood Estimates
Model: Censored Normal (cnorm)

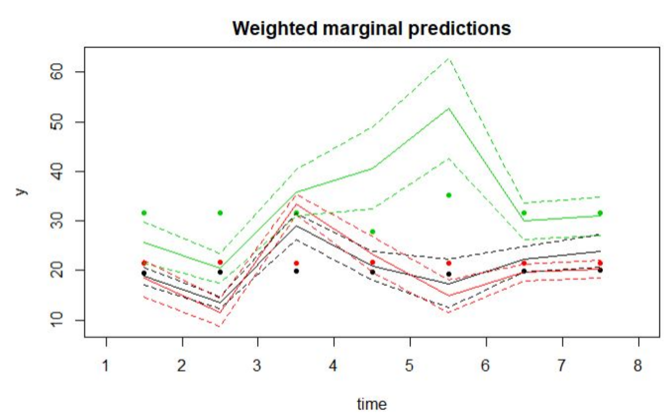
Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	22.82180	1.54163	14.804	0.00
	Linear	-0.78228	0.52952	-1.477	0.14
	Quadratic	0.13017	0.05742	2.267	0.02
	nd0	-0.35963	0.05018	-7.167	0.00
2	Intercept	31.75633	0.91769	34.605	0.00
	Linear	-0.87823	0.37941	-2.315	0.02
	Quadratic	0.15974	0.04120	3.877	0.00
	nd0	-0.49078	0.02229	-22.021	0.00
3	Intercept	38.67268	1.36874	28.254	0.00
	Linear	-1.89856	0.65560	-2.896	0.00
	Quadratic	0.33348	0.07116	4.687	0.00
	nd0	-0.38084	0.03320	-11.470	0.00
	Sigma	2.86032	0.09870	28.980	0.00
Group membership					
1	(%)	27.81198	6.30127	4.414	0.00
2	(%)	54.08856	6.98952	7.739	0.00
3	(%)	18.09946	5.29777	3.416	0.00

BIC= -1185.76 (N=440) BIC= -1170.17 (N=55) AIC= -1155.11 ll= -1140.11




```
Goodness-of-fit statistics:
     maximum log-likelihood: -872.75
     AIC: 1775.51
     BIC: 1805.62


Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                         coef      Se    Wald p-value
intercept class1    -0.97856 1.39558 -0.701 0.48319
intercept class2     1.22476 1.15093  1.064 0.28726
nodipl_rate class1   0.12580 0.06807  1.848 0.06458
nodipl_rate class2   0.02441 0.06300  0.388 0.69837

Fixed effects in the longitudinal model:

                       coef      Se   Wald p-value
intercept class1 17.88939 1.70334 10.503 0.00000
intercept class2 22.43568 1.64973 13.600 0.00000
intercept class3 29.42725 3.49491  8.420 0.00000
time class1      -0.53790 0.18496 -2.908 0.00364
time class2      -1.20038 0.17582 -6.827 0.00000
time class3      -1.81050 0.38560 -4.695 0.00000
I(time^2) class1  0.16402 0.01989  8.245 0.00000
I(time^2) class2  0.17580 0.01962  8.961 0.00000
I(time^2) class3  0.40405 0.04320  9.353 0.00000
```

~~~
```{r}
plot(mod_6, which='fit',var.time='time',legend=NULL)
```
~~~



Model 6: three clusters with quadratic mean and percent of population below poverty rate

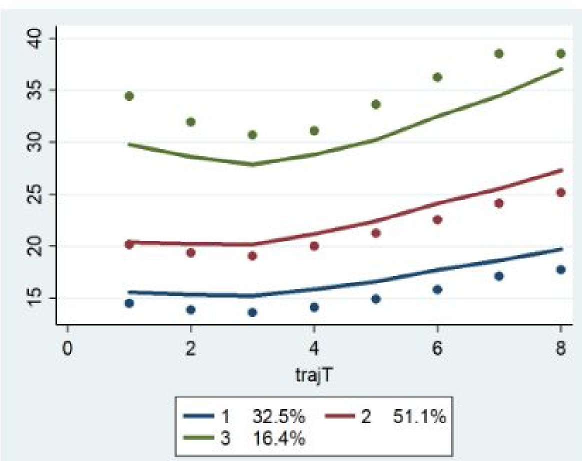
```
%%stata
use "final_1_wide.dta", clear
traj , model(cnorm) var(Y2010 Y2011 Y2012 Y2013 Y2014 Y2015 Y2016 Y2017) indep(X1-X8) order(2 2 2) tcov(bp0 bp1 bp2 bp3 bp4 bp5 bp6 bp7) min(-100) max(100)
trajplot
```

```
mod_8 <- hlme(diversion_rate_total~time+I(time^2),
              mixture=~time+I(time^2),
              classmb=~below_pover, random=~1,subject='puma',ng=3,
              data=data_long)
```

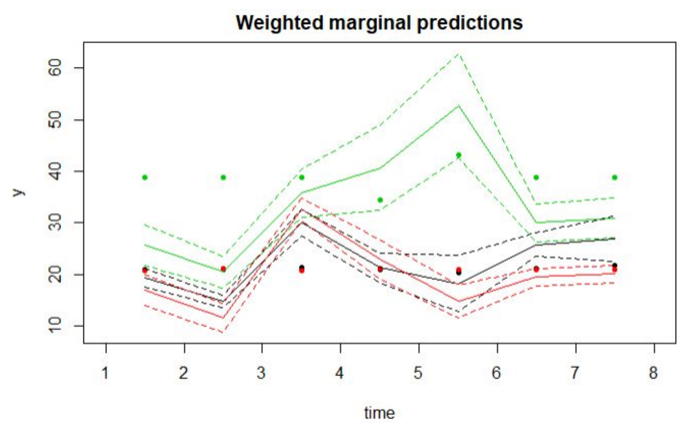
Maximum Likelihood Estimates
Model: Censored Normal (cnorm)

Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	21.76004	1.12314	19.374	0.00
	Linear	-0.34986	0.46763	-0.748	0.45
	Quadratic	0.10004	0.05075	1.971	0.04
	bp0	-0.33390	0.03104	-10.757	0.00
2	Intercept	28.74531	0.80732	35.606	0.00
	Linear	-0.36456	0.37259	-0.978	0.32
	Quadratic	0.14373	0.04040	3.558	0.00
	bp0	-0.45778	0.01880	-24.356	0.00
3	Intercept	43.37475	1.47077	29.491	0.00
	Linear	-1.60613	0.66281	-2.423	0.01
	Quadratic	0.28457	0.07190	3.958	0.00
	bp0	-0.69179	0.06575	-10.521	0.00
	Sigma	2.72694	0.09428	28.923	0.00
Group membership					
1	(%)	32.50344	6.52029	4.985	0.00
2	(%)	51.08300	6.95543	7.344	0.00
3	(%)	16.41356	5.09099	3.224	0.00

BIC= -1165.20 (N=440) BIC= -1149.60 (N=55) AIC= -1134.55 ll= -1119.55




```
Goodness-of-fit statistics:
     maximum log-likelihood: -869.1
     AIC: 1768.19
     BIC: 1798.3


Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                       coef      Se    Wald p-value
intercept class1   -2.66669 2.32633 -1.146 0.25167
intercept class2   -1.06330 2.24699 -0.473 0.63606
below_pover class1  0.41417 0.24875  1.665 0.09591
below_pover class2  0.33463 0.24659  1.357 0.17477

Fixed effects in the longitudinal model:

                     coef      Se    Wald p-value
intercept class1 18.96276 1.54812 12.249 0.00000
intercept class2 21.59995 1.52298 14.183 0.00000
intercept class3 37.96260 5.14543  7.378 0.00000
time class1      -0.58917 0.16876 -3.491 0.00048
time class2      -1.12508 0.16286 -6.908 0.00000
time class3      -3.19031 0.57127 -5.585 0.00000
I(time^2) class1  0.18076 0.01824  9.908 0.00000
I(time^2) class2  0.16988 0.01762  9.644 0.00000
I(time^2) class3  0.59512 0.06226  9.559 0.00000
```

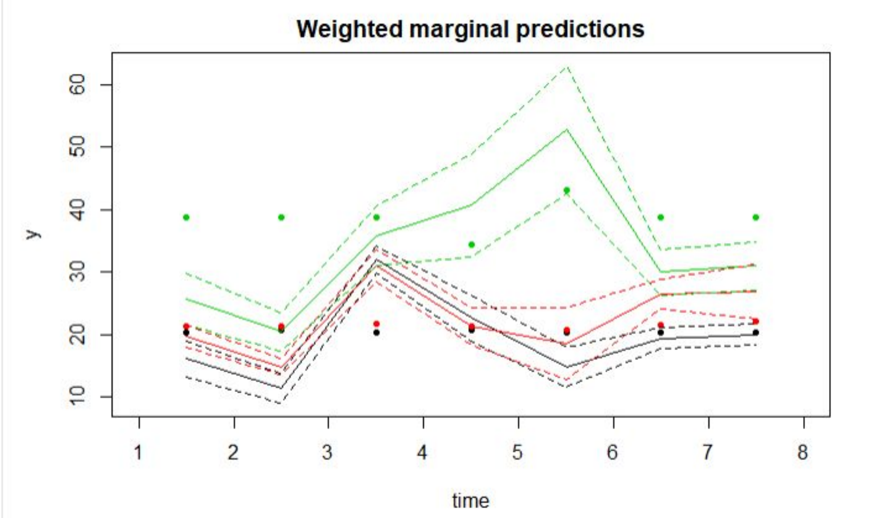


The addition of the time dependent predictor to the traj model decreased BIC, while all intercepts and quadratic terms are significant. While all coefficients produced by the hlme model are significant and BIC smaller than in the previous models, the overlapping clusters depicted on the graph suggest that the two overlapping clusters might be just one cluster.

Model Profiles

a) hlme model with three clusters and quadratic terms

```{r}
plot(mod_4, which='fit',var.time='time',legend=NULL)
```

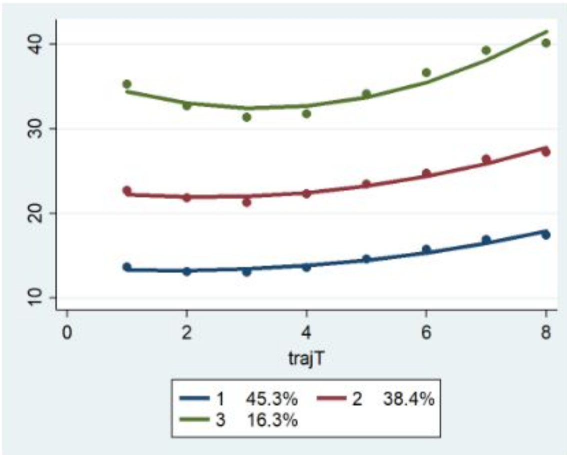


class	Solid waste diversion rate	Average percent of population with no high school diploma, 2017	Average percent of population below poverty rate, 2017
3	**49.94**	**10.09**	**8.93**
2	**25.73**	**21.40**	**22.27**
1	**22.51**	**17.76**	**17.94**

class	borough	count
1	Bronx	3
1	Brooklyn	11
1	Manhattan	4
1	Queens	9
1	Staten Island	2
2	Bronx	7
2	Brooklyn	6
2	Manhattan	6
2	Queens	4
2	Staten Island	1
3	Brooklyn	1
3	Queens	1

The analysis of the cluster composition confirmed that the three-cluster hlme model with quadratic term was a two cluster group with just a few diverging sub-areas in Brooklyn and Queens in more recent years. These results are insufficient to suggest that they represent an independent cluster.

b) Traj quadratic model with three clusters



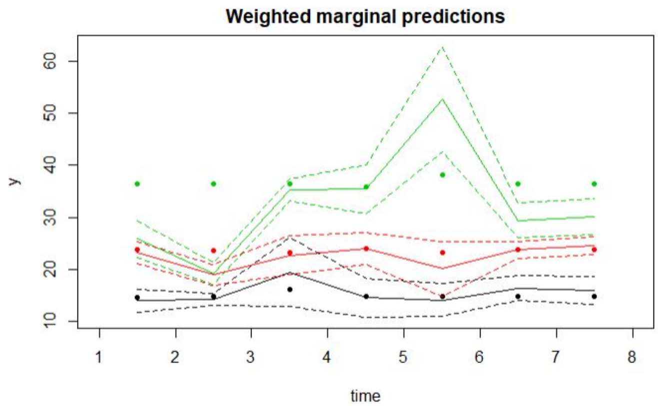
class	Solid waste diversion rate	Average percent of population with no high school diploma, 2017	Average percent of population below poverty rate, 2017
3	40.16	7.46	10.51
2	27.26	17.87	15.15
1	17.46	24.25	26.39

_traj_Group	borough	count
1	Bronx	7
1	Brooklyn	11
1	Manhattan	3
1	Queens	4
2	Bronx	3
2	Brooklyn	4
2	Manhattan	2
2	Queens	9
2	Staten Island	3
3	Brooklyn	3
3	Manhattan	5
3	Queens	1

The Traj quadratic model with three clusters presents a better defined set of clusters.

c) Hlme three cluster linear model

```{r}
plot(mod_2,which='fit',var.time='time',legend=NULL)
```

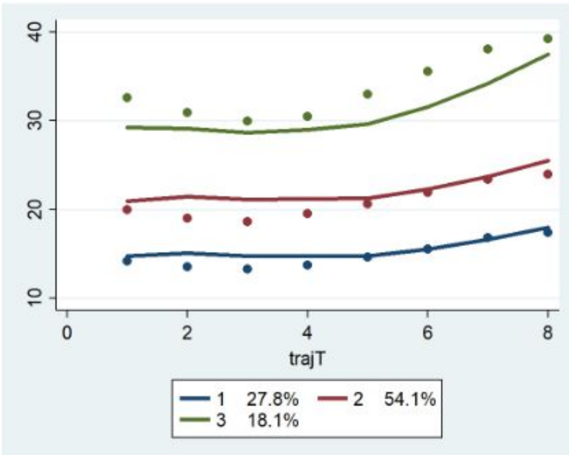


class	Solid waste diversion rate	Average percent of population with no high school diploma, 2017	Average percent of population below poverty raet, 2017
3	40.35	5.43	10.10
2	27.43	17.98	15.56
1	17.49	24.41	26.65

class	borough	count
1	Bronx	6
1	Brooklyn	9
1	Manhattan	4
1	Queens	4
2	Bronx	4
2	Brooklyn	7
2	Manhattan	1
2	Queens	10
2	Staten Island	3
3	Brooklyn	2
3	Manhattan	5

The hlme linear three-cluster model produced results that are very similar to the traj quadratic model with three clusters.

d) Traj three cluster model with quadratic mean and a time-dependent predictor (percent of population with no high school diploma)



class	Solid waste diversion rate	Average percent of population with no high school diploma, 2017	Average percent of population below poverty rate, 2017
3	39.21	11.81	12.14
2	23.91	21.3	20.17
1	17.46	19.49	23.07

_traj_Group	borough	count
1	Bronx	2
1	Brooklyn	10
1	Manhattan	1
1	Queens	2
2	Bronx	8
2	Brooklyn	4
2	Manhattan	5
2	Queens	10
2	Staten Island	3
3	Brooklyn	4
3	Manhattan	4
3	Queens	2

The addition of the time-varying covariate of population with no high school diploma somewhat altered the composition of each cluster.

Results

Based on the BIC values and graphical analysis the Traj three cluster model with quadratic mean and the Traj three cluster model with quadratic mean and a time-dependent predictor produced the most well-defined results. However, it should be noted that these two models are not the same, as with additional variables cluster composition changed.

Results indicated that there are three groups of municipal sub-areas in New York that pursued similar trajectories with respect to solid waste diversion rates over 2010-2017. The profiles based on the clustering data and demographic characteristics suggest that trajectory clustering is generally consistent with the theory that high diversion rates are associated with higher levels of education and lower levels of poverty, except for the model with an additional time-dependent predictor: a cluster with lower levels of education has higher levels of recycling. The application of clusters to the diversion rate data for this last model is presented below.