Milestone 8: To Turn In

Maria Burzillo

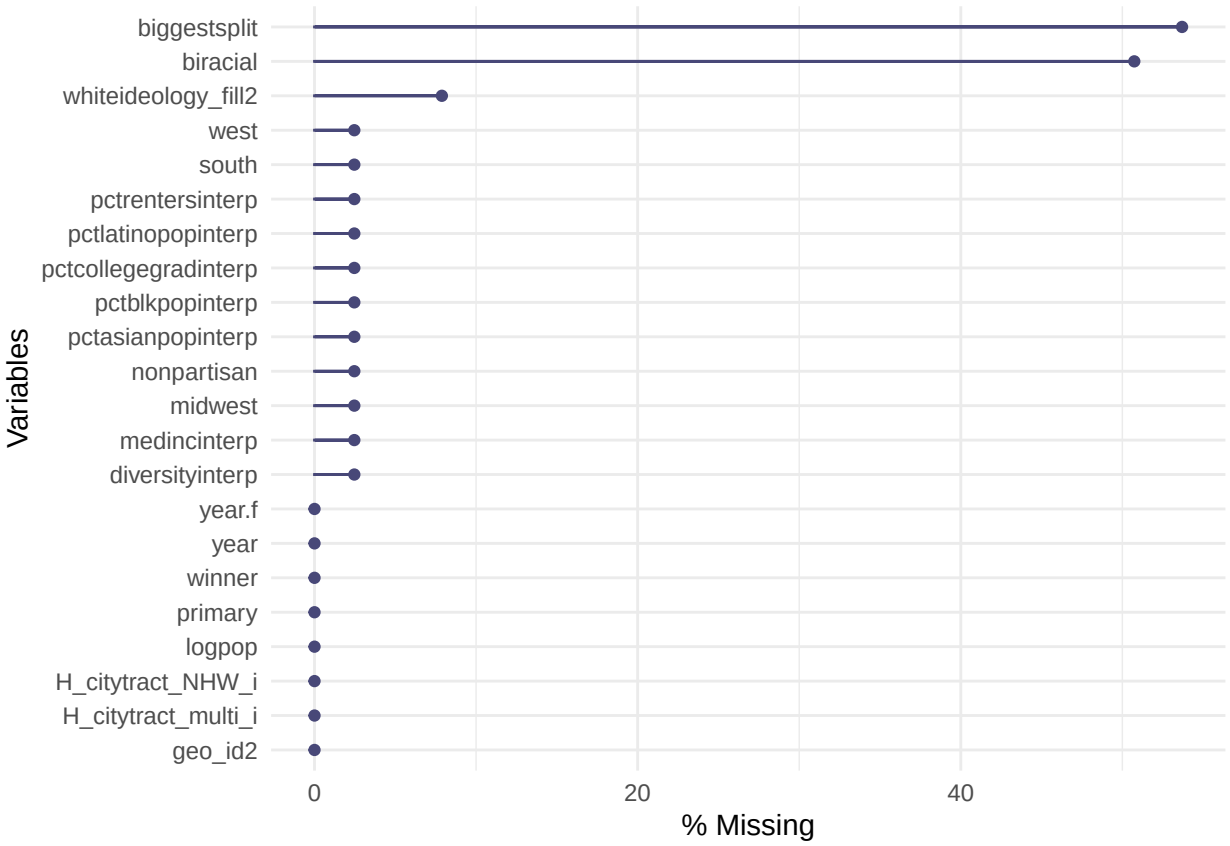
3/28/2020

Abstract

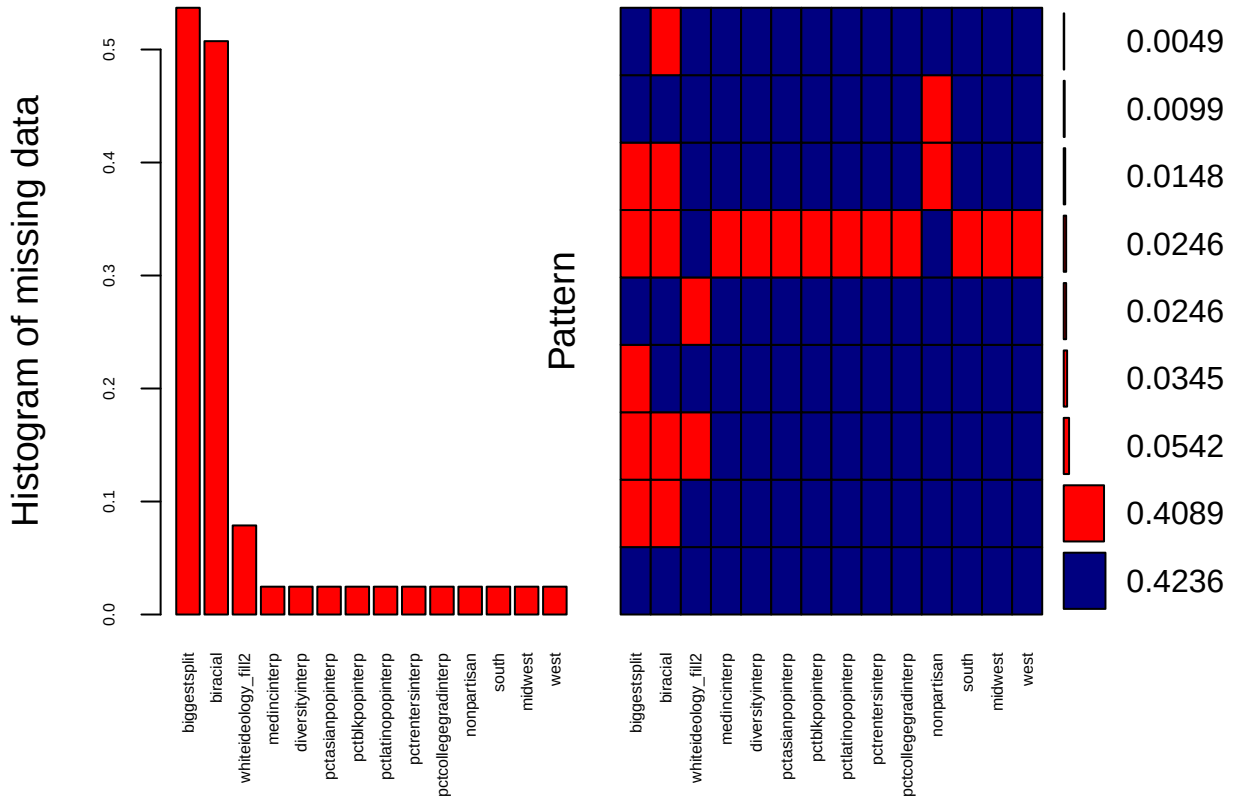
This is an extension of Jessica Trounstine's "Segregation and Inequality in Public Goods" (2016). I was able to replicate the main results of Trounstine's paper in R to suggest that racial segregation contributes to political polarization and decreased spending on public goods. Additionally, I extend the analysis by imputing missing data and rerunning Trounstine's original model as a robustness check.

Extension 1

One aspect of Trounstine's paper with room for improvement is that there is a large amount of missing data in her datasets upon which she bases her analyses. Using R's mice package (add citation), we can perform multiple imputation. By performing imputation multiple times, this helps account for the uncertainty inherent in the individual imputations. Before performing the multiple imputations, we will first look at the missing data to see if there are any patterns.



The above plot reveals that biggestsplit and biracial are the two variables with by far the largest percentage of missing values. Both of these are missing approximately 50% of their observations. Since biggestsplit is the dependent variable in our analysis with this dataset, this is an important fact.

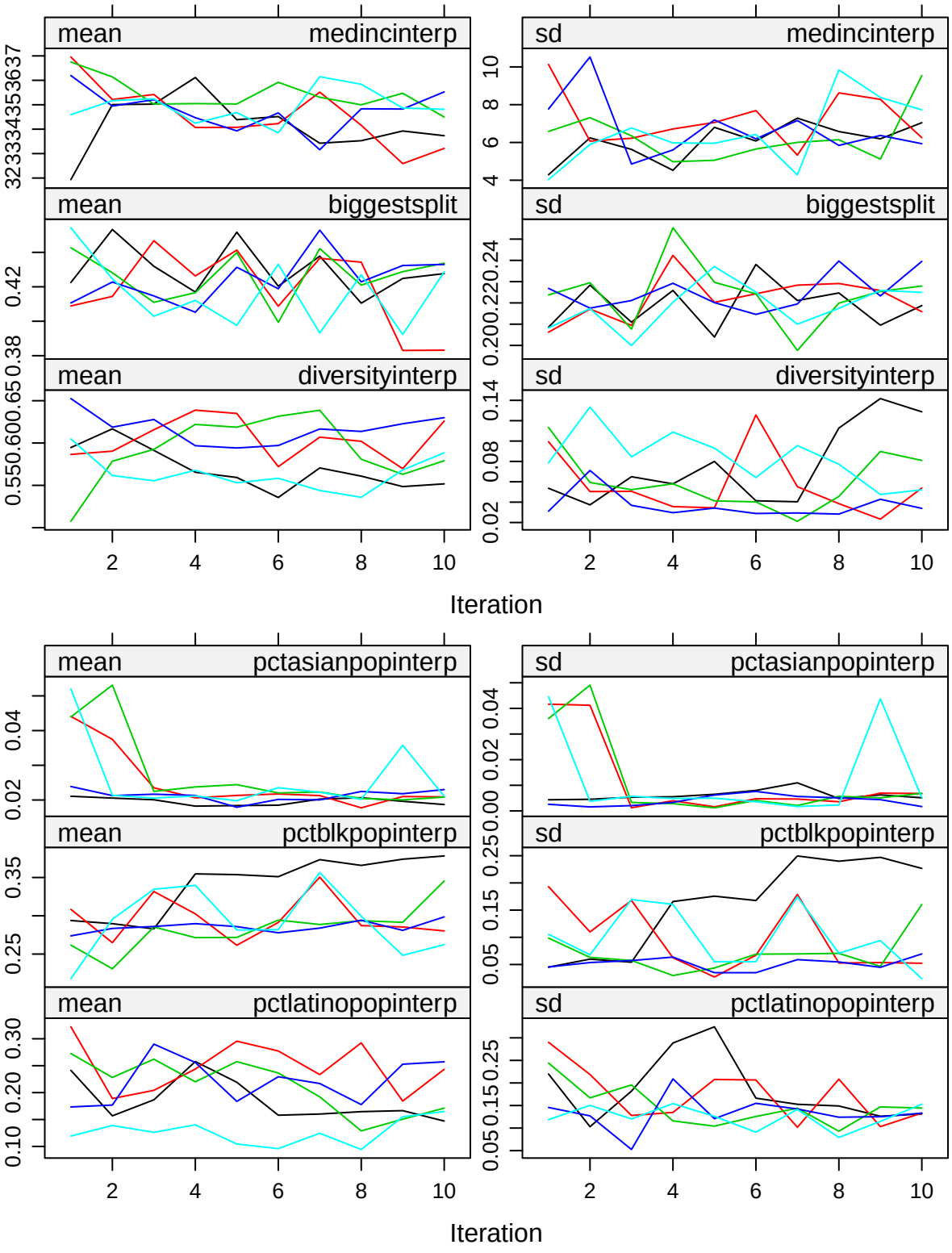


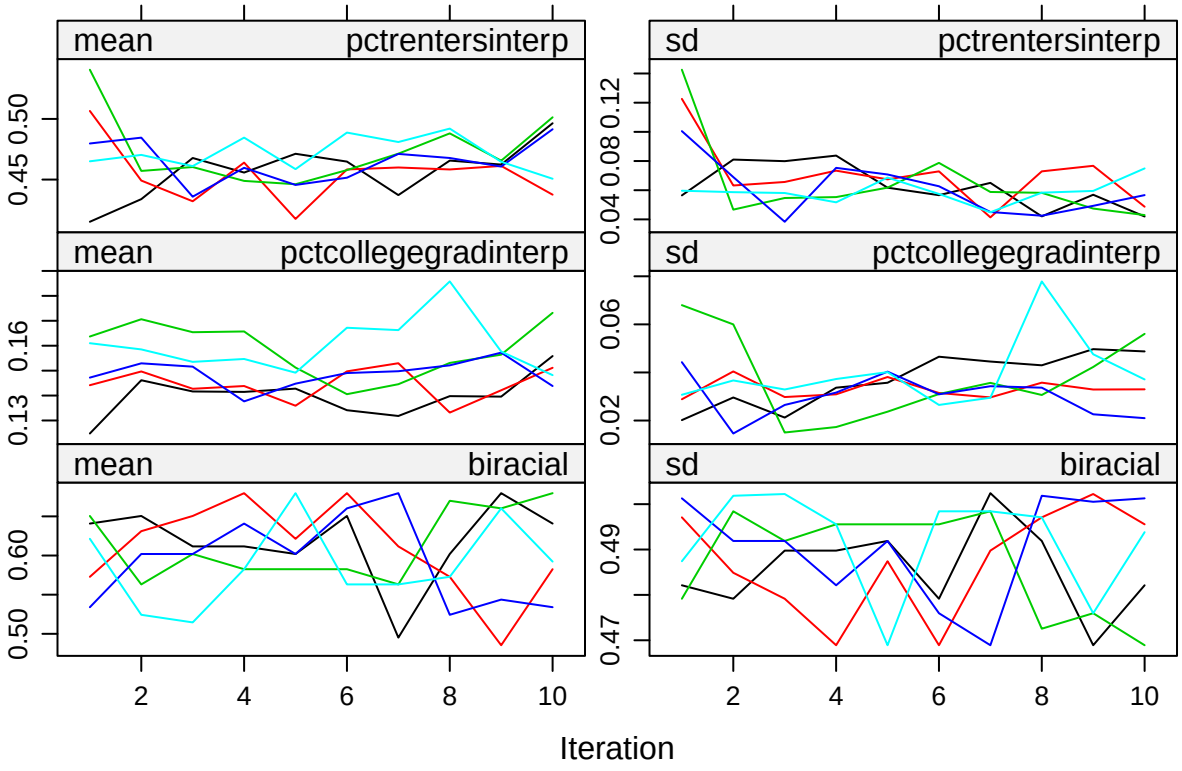

```
Variables sorted by number of missings:
            Variable      Count
         biggestsplit 0.53694581
             biracial 0.50738916
  whiteideology_fill2 0.07881773
        medincinterp 0.02463054
     diversityinterp 0.02463054
   pctasianpopinterp 0.02463054
     pctblkpopinterp 0.02463054
  pctlatinopopinterp 0.02463054
    pctrentersinterp 0.02463054
pctcollegegradinterp 0.02463054
          nonpartisan 0.02463054
                south 0.02463054
              midwest 0.02463054
                 west 0.02463054
```

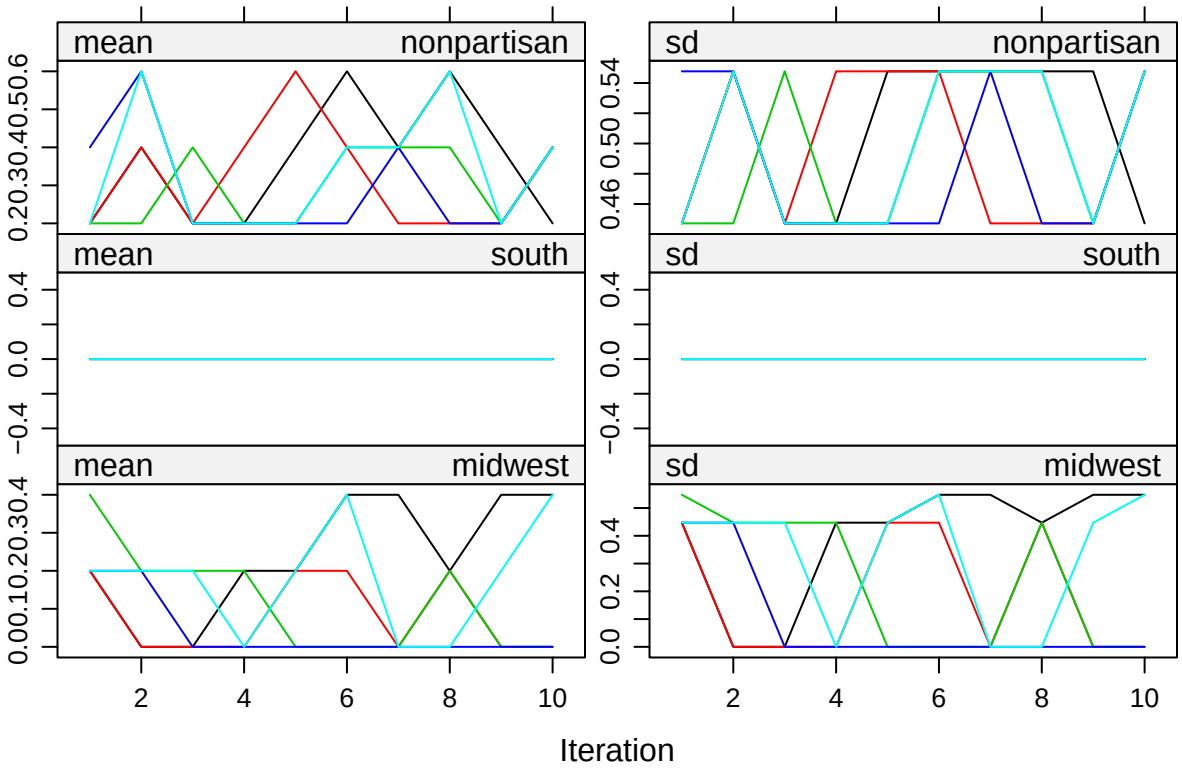
To better understand any potential patterns in missing data, we then plotted the pattern of missingness for only those variables missing values. From the plot on the right, we can see that approximately 42% of observations are complete. There seem to be a correspondence between missing a value for biggest split and missing biracial. There also seem to be about 2% of values for wich most of the variables are missing. However, most observations are not missing more than 2-3 values.

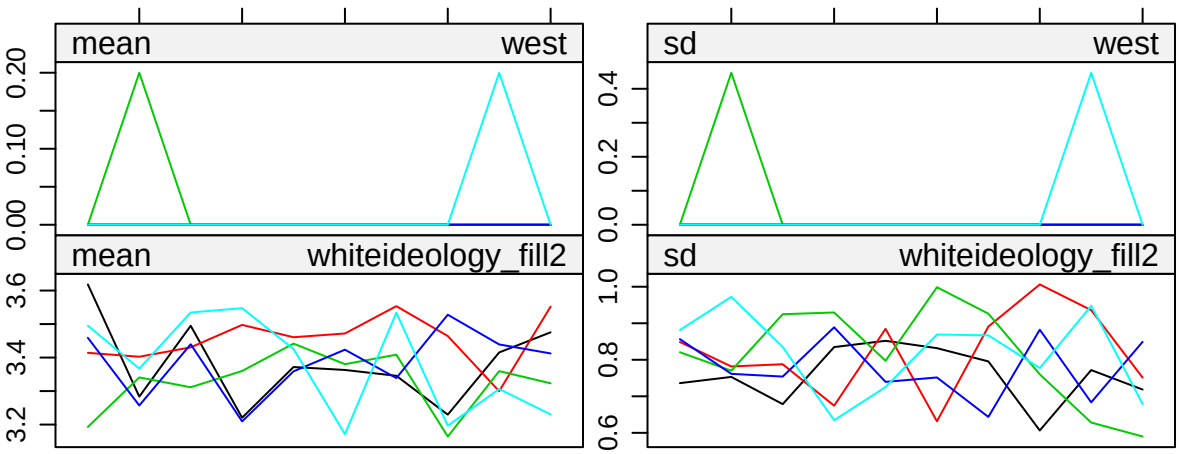
Finding no clear patterns in the missing data, I next performed mutliple imputations (with 10 iterations) on the dataset. A non-stochastic imputation method, Classification and Regression Trees (CART), was used instead of the default because of an error with matrix inversion caused by the data. Before examining the results of Trounstine's model using the imputed data, I first run some diagnostic tests of the imputation results to make sure that everything is running as expected.

First, I check the convergence of the algorithm used within mice() for each of the variables. For the most part, the fits intertwine and do not exhibit any trends at later iterations, as desired. There are some issues with the results for some of the variables with very few missing values (such as West and South), which makes sense as we would expect the mean to be less reliable due to the law of large numbers, and convergence is more difficult. Nevertheless, since there are so few of these values in the actual dataset, this is not a major concern.



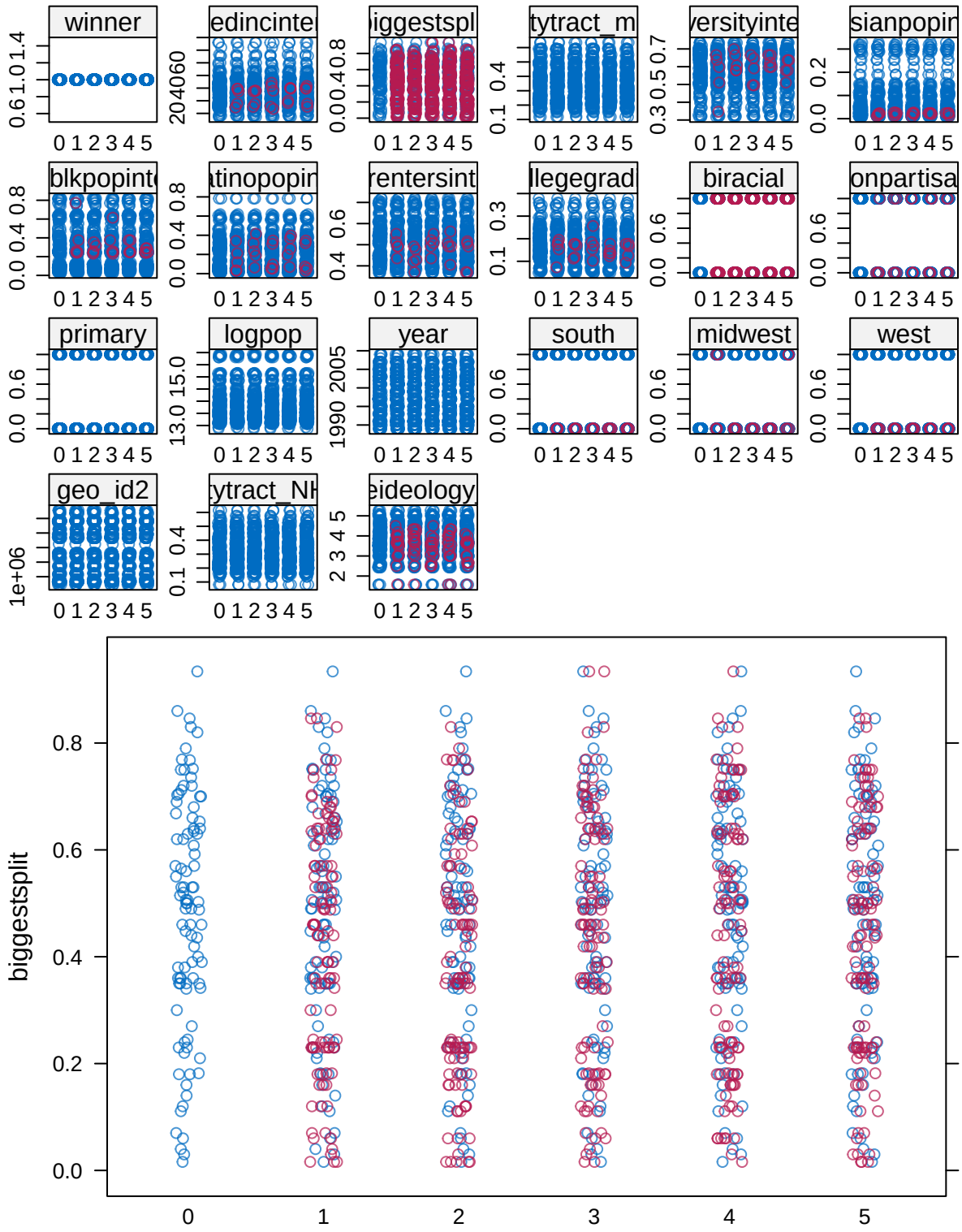

mean
pctrentersinterp
sd
pctrentersinterp
mean
pctcollegegradinterp
sd
pctcollegegradinterp
mean
biracial
sd
biracial
Iteration


mean
nonpartisan
sd
nonpartisan
mean
south
sd
south
mean
midwest
sd
midwest
Iteration

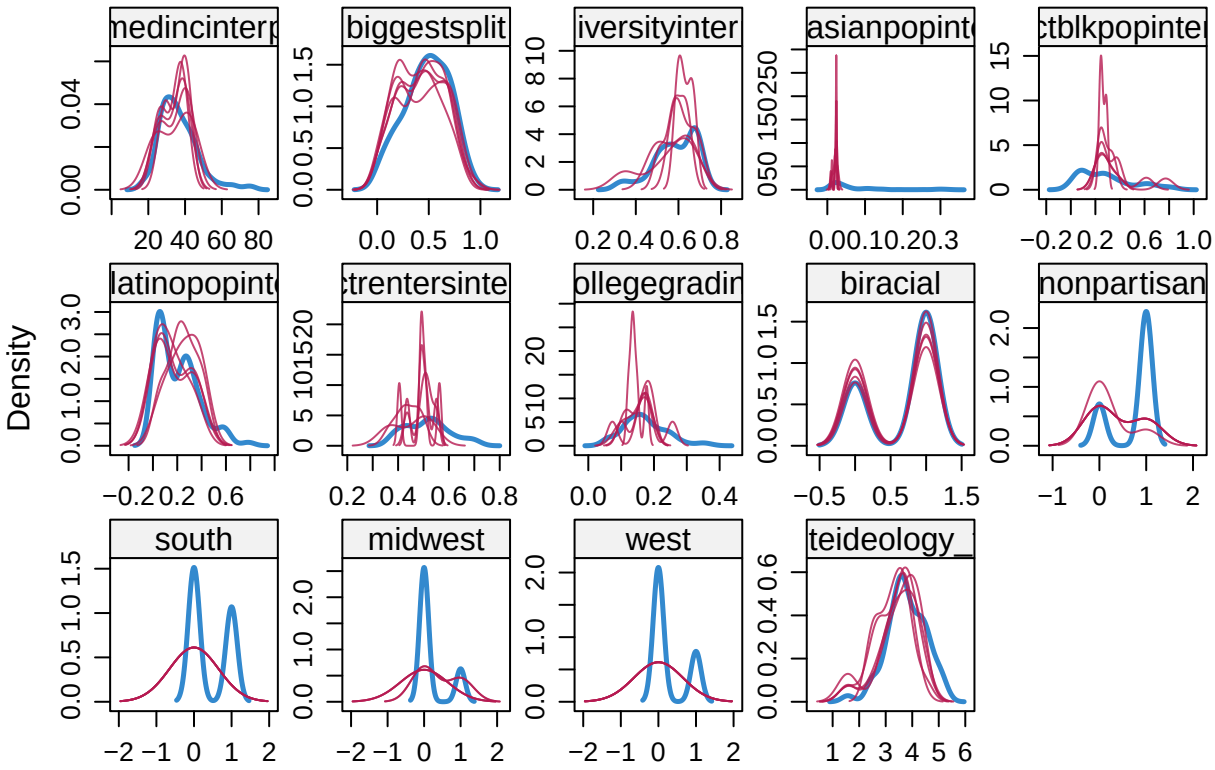


Iteration

We can also check the imputed values against the original values using stripplot(). Each column in each subplot represents a separate iteration. The magenta points represent the imputed data. The values of the variable in questions are along the y axis. We expect the spread of the data to be similar if the imputations were done well. If the data were missing completely at random, then the imputed data should have the same distribution as the original data. In particular, we want to be sure that the imputations are within a plausible range of the data. This is the case for all of our imputed variables, and there does not seem to be cause for alarm from these results.



Finally, we can also look at the density plots for each variable's actual data and for their imputed data from each of the iterations, which are represented in magenta. Overall, the density plots align quite well for the variables with the most missing data, biracial and biggestsplit, and relatively well for most of the other variables with missing data. Again, the fit is less good for variables with fewer missing data points.



While the imputations are not perfect, there do not seem to be any major problems so far. Thus, we can now proceed with our analysis. We fit each of our ten imputed datasets to Trounstine's 3 models using this data and then pool the results for each. The results are as follows:

Extension 1 Racial Polarization Table

Regression 1

	term	estimate	std.error	p.value
1	(Intercept)	-0.0777729776	0.36517260	8.318992e-01
2	H_citytract_multi_i	0.5002192474	0.37548390	2.060719e-01
3	diversityinterp	0.1111479459	0.36294671	7.644052e-01
4	pctasianpopinterp	-0.2618726727	0.37444724	4.876252e-01
5	pctblkpopinterp	0.1110694606	0.19214994	5.683679e-01
6	pctlatinopopinterp	0.0278772514	0.19961536	8.904540e-01
7	medincinterp	0.0002671111	0.00425506	9.504653e-01
8	pctrentersinterp	-0.3827322963	0.35697040	3.074446e-01
9	pctcollegegradinterp	0.3017655854	0.72557472	6.894317e-01
10	biracial	0.2198877085	0.03143064	2.417454e-07
11	nonpartisan	-0.0350971186	0.06413681	5.949594e-01
12	primary	-0.0610973565	0.02778458	3.747989e-02
13	logpop	0.0234278991	0.02924786	4.252126e-01

The results of the model with the new dataset are slightly different in comparison to the original results from Trounstine (2016). With the imputed data, we now have a total of 203 observations in our model as compared to the original model, which had only 91 observations. Interestingly, while the sign of the coefficient on the main variable of interest, the Theil's H segregation index is the same and the standard error has decreased slightly, the result has become statisically insignificant and the 95% confidence interval contains zero. Thus while these results still suggest that segregation may be associated with increased political polarization, they

confer a lesser degree of certainty than Trounstine's original analysis. The coefficients for pctblkpopinterp, pctlatinopopinterp, and medincinterp have also switched signs, and all coefficients except indicators for a biracial and primary election are also statistically insignificant, as was the case in the original analysis. In general, the standard error on the coefficients has decreased slightly.

In the second and third regressions, the results are similar in comparison to the unimputed data to the first. Standard errors have reduced slightly. The main coefficient on segregation index has decreased in magnitude, although in these specifications, it remains statistically significant. The other coefficients except for biracial and primary are insignificant, and the coefficients on percent Asian, Black, and Latino have all switched signs.

Regression 2

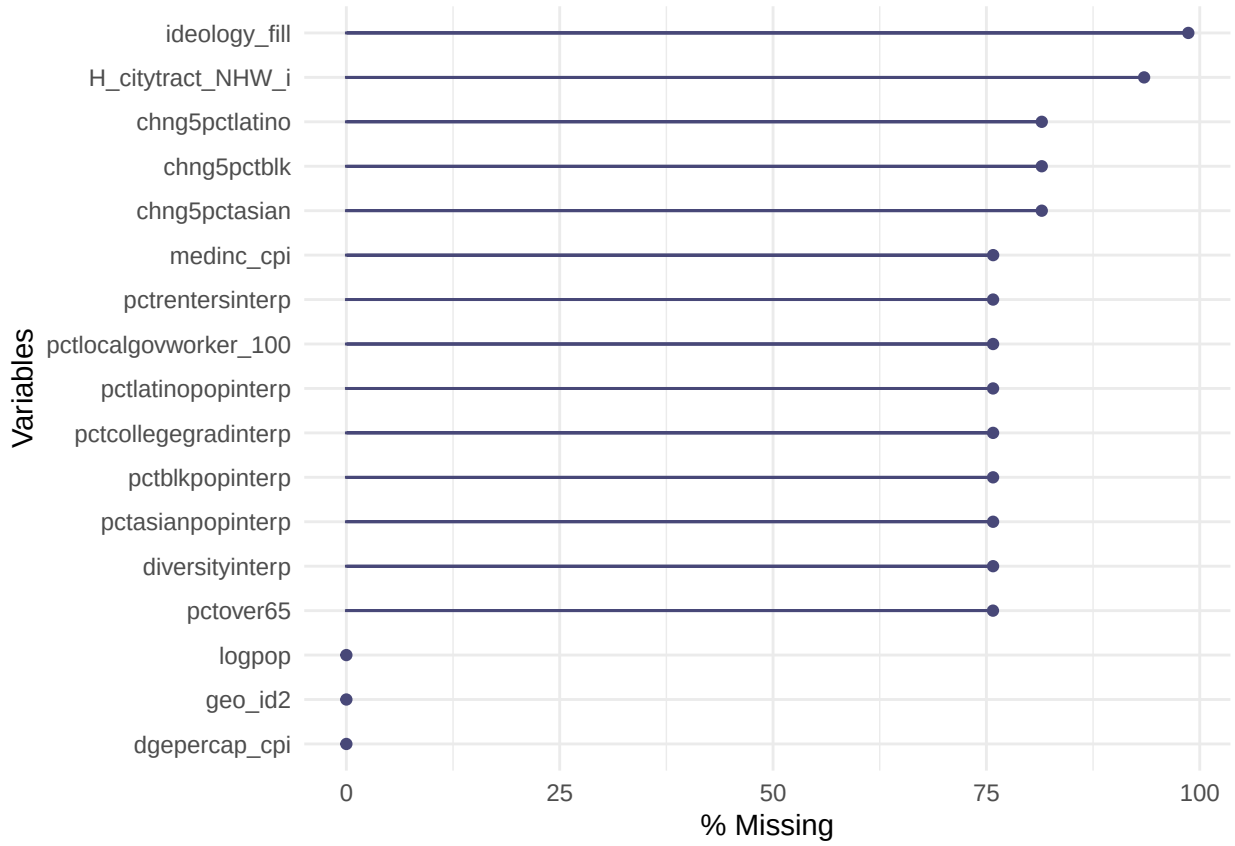
	term	estimate	std.error	p.value
1	(Intercept)	-0.1411965424	0.372535202	7.060703e-01
2	H_citytract_NHW_i	0.5986092142	0.274070125	4.083681e-02
3	diversityinterp	0.0695064864	0.303184361	8.212224e-01
4	pctasianpopinterp	-0.0599621545	0.389002316	8.781525e-01
5	pctblkpopinterp	0.1368823542	0.157090533	3.893361e-01
6	pctlatinopopinterp	0.0802965700	0.195690633	6.862202e-01
7	medincinterp	-0.0007028223	0.003992971	8.613522e-01
8	pctrentersinterp	-0.4766670633	0.353130552	2.038871e-01
9	pctcollegegradinterp	0.4749425107	0.716547045	5.273236e-01
10	biracial	0.2158444492	0.031734983	6.641788e-07
11	nonpartisan	-0.0199275151	0.061627255	7.517596e-01
12	primary	-0.0598519738	0.027364480	3.815474e-02
13	logpop	0.0274986554	0.028884262	3.435469e-01

Regression 3

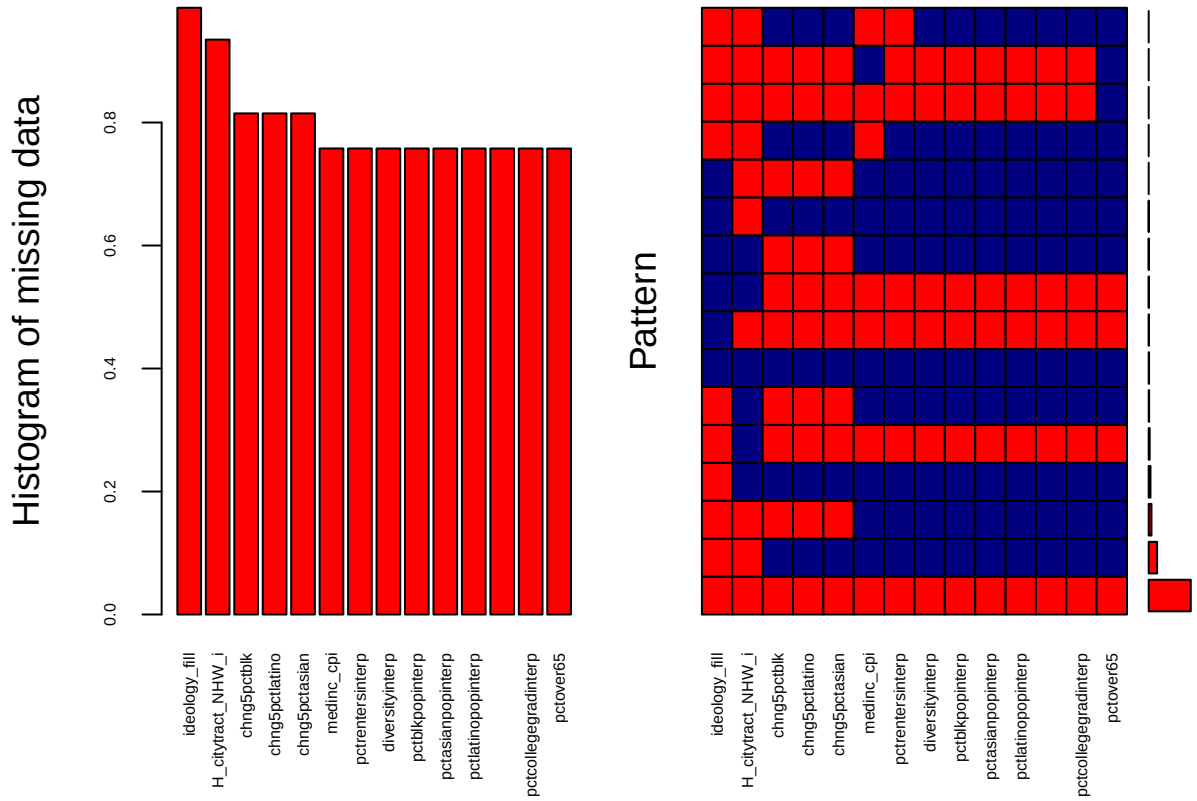
	term	estimate	std.error	p.value
1	(Intercept)	-0.1581525974	0.371888752	6.722244e-01
2	H_citytract_NHW_i	0.5794593451	0.274555897	4.749750e-02
3	diversityinterp	0.1026921123	0.311053972	7.454200e-01
4	pctasianpopinterp	-0.0388935699	0.392401865	9.215050e-01
5	pctblkpopinterp	0.1524045830	0.149545826	3.123959e-01
6	pctlatinopopinterp	0.0692509356	0.202699816	7.371024e-01
7	medincinterp	-0.0003156848	0.004011103	9.377374e-01
8	pctrentersinterp	-0.5096252757	0.347079682	1.677241e-01
9	pctcollegegradinterp	0.3787351759	0.786797599	6.457132e-01
10	biracial	0.2129184907	0.031828172	7.836879e-07
11	nonpartisan	-0.0153707293	0.061180115	8.055409e-01
12	primary	-0.0598942534	0.027397102	3.853823e-02
13	logpop	0.0331789273	0.030293379	2.771331e-01

Extension Part 2: Fin Seg Data Set

In the second, and arguably most important dataset in the analysis, the data set on financial segregation, there are a very large number of missing values. For example, on the main independent variable of interest, there are a whopping 282,334 missing values, or %87.09 of the data.




```
Warning in plot.aggr(res, ...): not enough horizontal space to display
frequencies
```

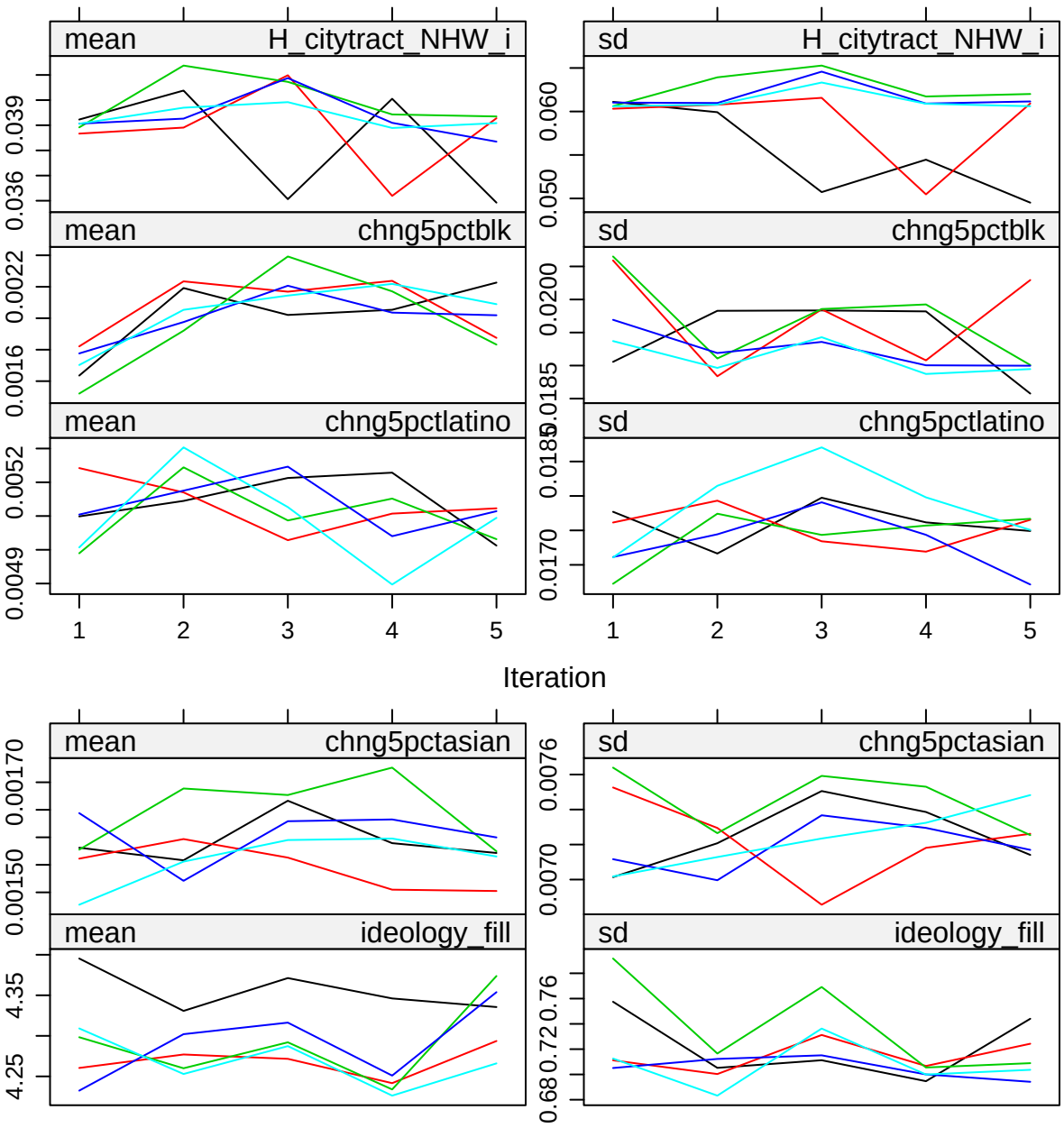



```
Variables sorted by number of missings:
             Variable     Count
         ideology_fill 0.9866041
    H_citytract_NHW_i 0.9347787
          chng5pctblk 0.8148547
       chng5pctlatino 0.8148547
        chng5pctasian 0.8148547
           medinc_cpi 0.7578741
     pctrentersinterp 0.7577980
      diversityinterp 0.7577914
      pctblkpopinterp 0.7577914
    pctasianpopinterp 0.7577914
   pctlatinopopinterp 0.7577914
pctlocalgovworker_100 0.7577914
  pctcollegegradinterp 0.7577914
             pctover65 0.7576722
```

```
[1] 1751.791
```

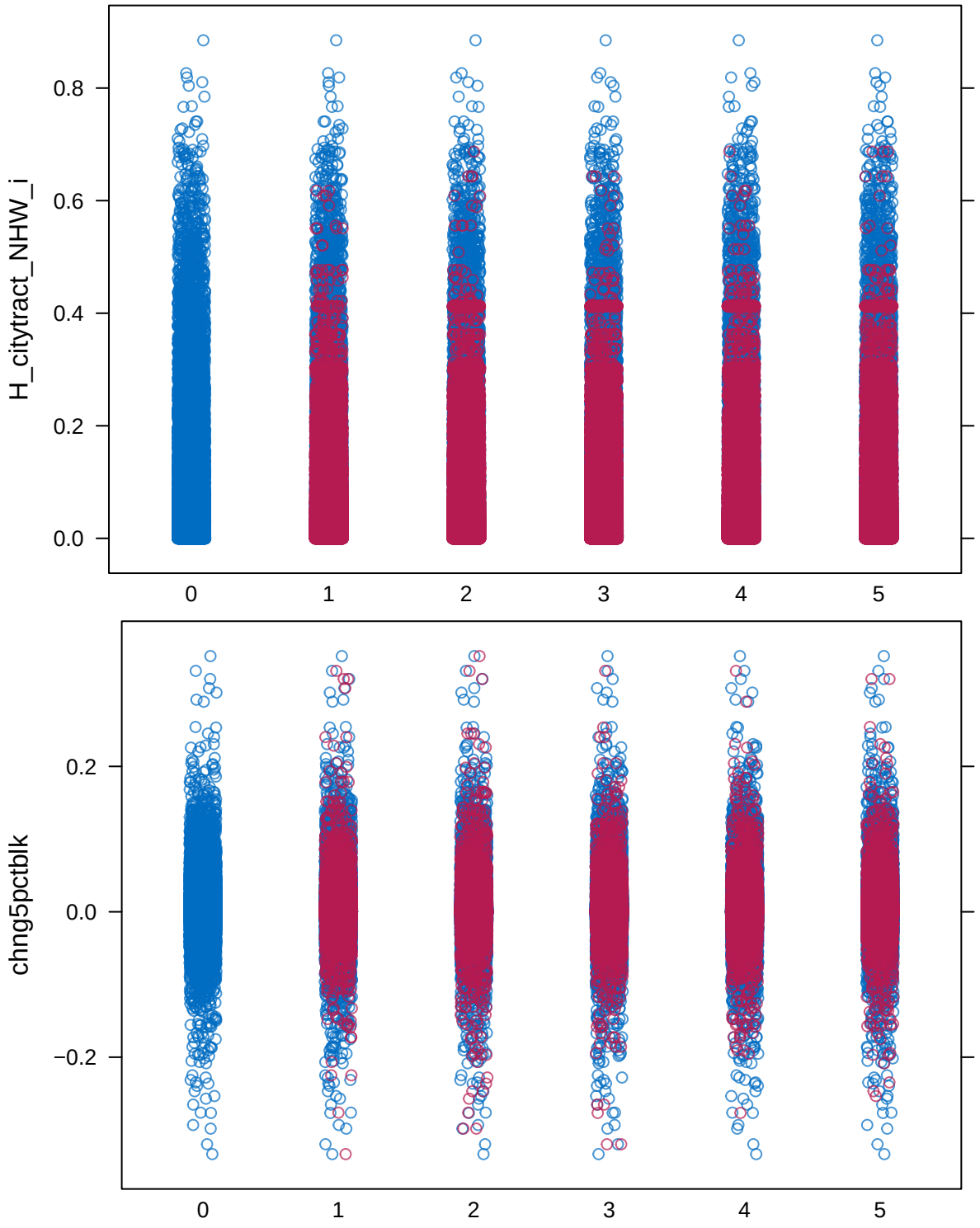
Finding no clear patterns in the missing data, I next performed mutliple imputations (with 5 iterations) on the dataset. A non-stochastic imputation method, Classification and Regression Trees (CART), was used instead of the default because of an error with matrix inversion caused by the data. Before examining the results of Trounstine's model using the imputed data, I first run some diagnostic tests of the imputation results to make sure that everything is running as expected.

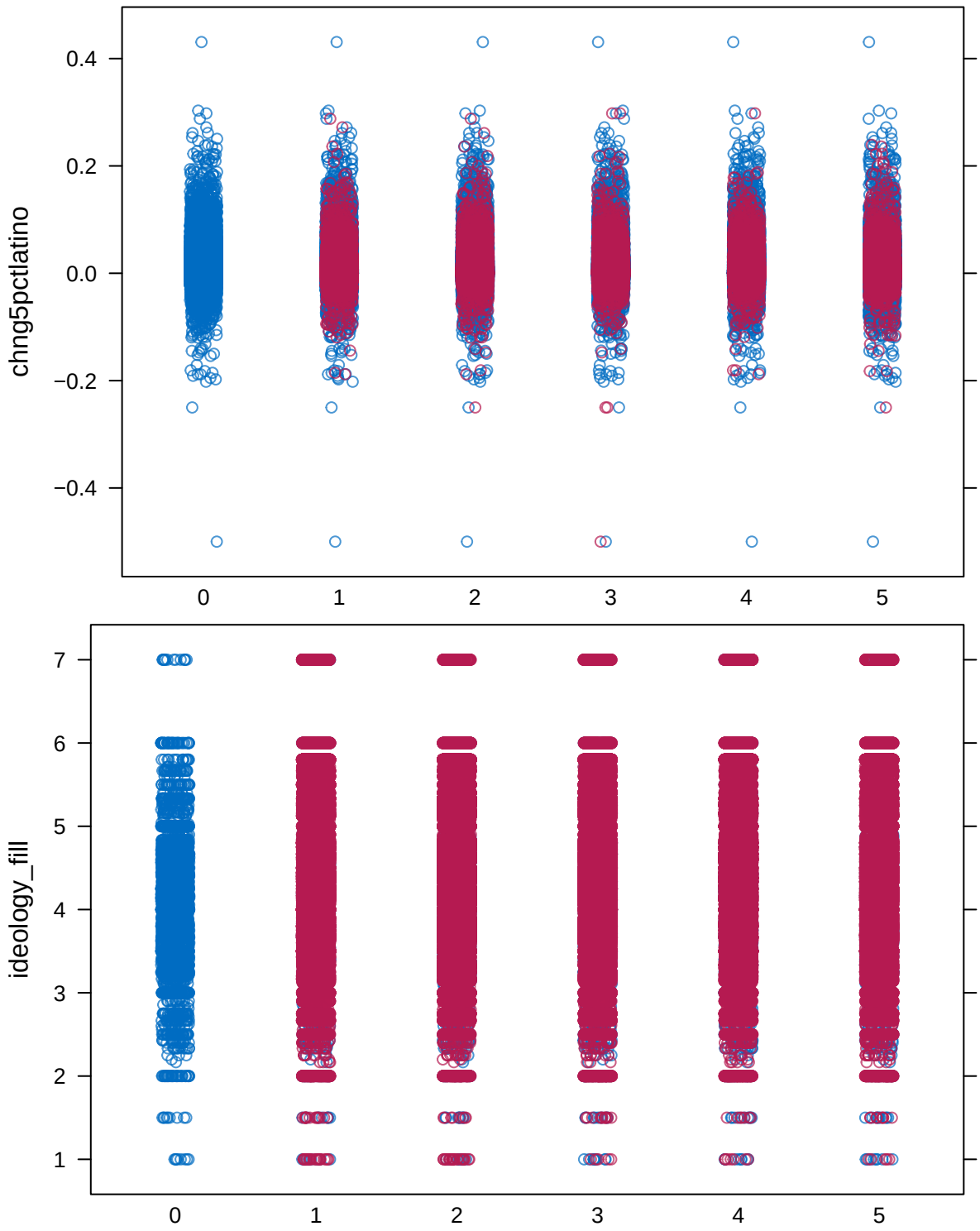
First, I check the convergence of the algorithm used within mice() for each of the variables. For the most part, the fits intertwine and do not exhibit any trends at later iterations, as desired.



We can also check the imputed values against the original values using stripplot(). Each column in each subplot represents a separate iteration. The magenta points represent the imputed data. The values of the variable in questions are along the y axis. We expect the spread of the data to be similar if the imputations were done well. If the data were missing completely at random, then the imputed data should have the same distribution as the original data. In particular, we want to be sure that the imputations are within a plausible

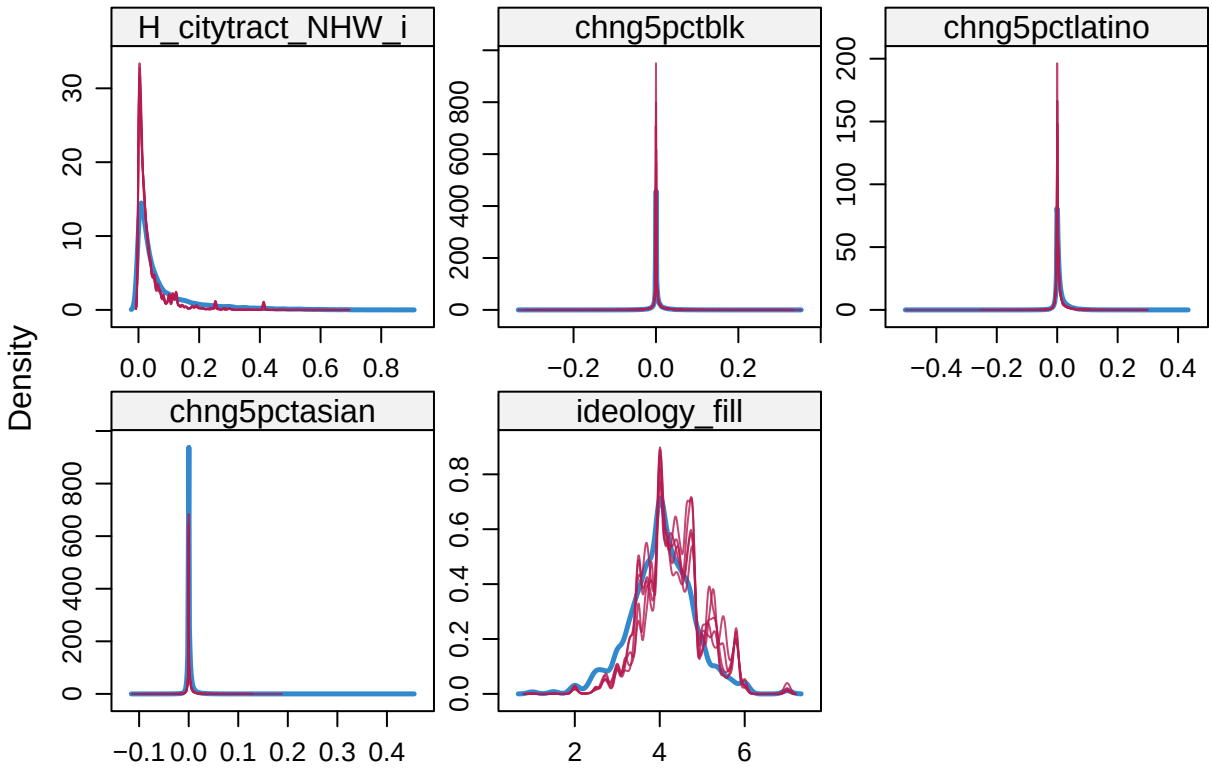
range of the data. This is the case for all of our imputed variables, and there does not seem to be cause for alarm from these results.





Finally, we can also look at the density plots for each variable's actual data and for their imputed data from each of the iterations, which are represented in magenta. Overall, the density plots align quite well for the variables.

```
Error in density.default(x = c(NA_real_, NA_real_, NA_real_, NA_real_, : need at least 2 points to sele
```



While the imputations are not perfect, there do not seem to be any major problems so far. Thus, we can now proceed with our analysis. We fit each of our 5 imputed datasets to Trounstine's 3 models using this data and then pool the results for each. The results are as follows:

Regression 1

The results of the first regression indicate that Trounstine's results are robust to the inclusion of the imputed data, which led to the inclusion of an additional 59,377 observaitons in the analysis. The coefficients, standard errors, and significance levels are essentially unchanged. The same is largely true for the second and third specifications, but the magnitude of the coefficient on the segregation decreases by a marginal amount.

	term	estimate	std.error	p.value
1	H_citytract_NHW_i	0.139725256	0.405718450	7.354049e-01
2	diversityinterp	0.751272157	0.292778007	1.029611e-02
3	pctblkpopinterp	0.483301142	0.349350843	1.665525e-01
4	pctasianpopinterp	0.982994100	1.262961669	4.363882e-01
5	pctlatinopopinterp	1.581181142	0.437406977	3.013220e-04
6	medinc_cpi	0.005643113	0.006681168	3.983292e-01
7	pctlocalgovworker_100	0.031454113	0.035537814	3.761211e-01
8	pctrentersinterp	0.100817449	0.336111077	7.642168e-01
9	pctover65	0.607075256	0.775541302	4.337692e-01
10	pctcollegegradinterp	3.133200000	0.784707040	6.555782e-05
11	logpop	-0.588032056	0.146878066	6.265988e-05

	term	estimate	std.error	p.value
1	H_citytract_NHW_i	0.16503982	0.405131174	6.895458e-01
2	pctblkpopinterp	1.10322743	0.347221941	1.508047e-03
3	pctasianpopinterp	1.21463752	0.880731402	1.678961e-01
4	pctlatinopopinterp	2.43380932	0.464420046	1.707909e-07
5	chng5pctblk	-0.68189410	0.734200602	3.538030e-01

```
6         chng5pctlatino -0.77989334 0.767938892 3.123802e-01
7          chng5pctasian  2.20234074 2.041358302 2.809437e-01
8             medinc_cpi  0.00589425 0.006690293 3.783217e-01
9  pctlocalgovworker_100  0.03058461 0.035468413 3.885312e-01
10      pctrentersinterp  0.14369964 0.333230292 6.663049e-01
11             pctover65  0.53679998 0.774926779 4.885004e-01
12  pctcollegegradinterp  3.41945066 0.819127369 3.000849e-05
13               logpop -0.58243971 0.146434834 6.993432e-05
```

```
                   term     estimate   std.error      p.value
1      H_citytract_NHW_i  0.143024472 0.399318809 7.249954e-01
2        diversityinterp  0.751188208 0.292782206 1.030571e-02
3        pctblkpopinterp  0.488306559 0.348789993 1.615307e-01
4      pctasianpopinterp  0.978740038 1.259502848 4.371195e-01
5     pctlatinopopinterp  1.579793555 0.437055242 3.016405e-04
6             medinc_cpi  0.005646432 0.006682476 3.981442e-01
7  pctlocalgovworker_100  0.031321702 0.035553248 3.783407e-01
8       pctrentersinterp  0.084033205 0.337804997 8.035511e-01
9              pctover65  0.603482776 0.776052959 4.367972e-01
10  pctcollegegradinterp  3.133684615 0.784339933 6.487400e-05
11               logpop -0.588359060 0.146897573 6.221281e-05
12         ideology_fill -0.012129920 0.025925864 6.533201e-01
```