Data for Diplomas Sample Benchmark Analysis

Purpose

The purpose of this analysis was to determine the proportion of variance that district level cohort size and percent of low-income students accounted for in graduation rates. The sample consisted of 10,850 public school districts that reported Adjusted Cohort Graduation Rates (ACGR) in the 2011-12 school year. The ACGR is a method for tracking a group (or cohort) of students who enter high school together, as first-time 9th graders (or 10th graders, in schools that begin in 10th grade) and graduate "on-time" (i.e., within three or four years) with a regular high school diploma. The ACGR accounts (or adjusts) for students who transfer into a school, transfer to another school, or die. If the 2011-12 ACGR was reported as a range (e.g., 55-59), the median value was used in its place.

*Adjusted Cohort Graduation Rates: https://inventory.data.gov/dataset/73021f54-f5bf-4f34-86dd-5497ba4bdcfc/resource/32b21f42-b80b-4b1b-ad50-b361c5cecd99

Benchmark I Results

	Estimate	Std. Error	t value	Pr(>|t|)	
STNAMWASHINGTON	-7.48719	1.86391	-4.017	5.93e-05	***
STNAMWEST VIRGINIA	3.85928	2.77440	1.391	0.164244	
STNAMWISCONSIN	0.05236	1.75977	0.030	0.976262	
STNAMWYOMING	-9.03756	2.91598	-3.099	0.001944	**
PLI	-32.61625	0.78409	-41.598	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.29 on 11225 degrees of freedom
Multiple R-squared: 0.2481, Adjusted R-squared: 0.2449
F-statistic: 77.16 on 48 and 11225 DF, p-value: < 2.2e-16

In our first attempt at accounting for variation in high school graduation rates, we were able to improve our predictive power. After adding in a few parameters to our model (state name,PLI (percentage low income), our R-squared value was nearly 25%. Because there are too many states to include in the output sample above, we have trimmed our results down to show only the last few states for conceptualization. Note that some states have a positive coefficient of determination, and some have a negative. This supports the hypothesis that high school graduation rates vary by state.
*PLI = (ECD_COHORT_1112/ALL_COHORT_1112)

Benchmark II Results

	Estimate	Std. Error	t value	Pr(>|t|)	
STNAMWISCONSIN	-0.9668	1.6145	-0.599	0.549305	
STNAMWYOMING	-6.4714	2.6659	-2.427	0.015219	*
NATIVEAMERICAN_PERC	103.0951	4.2939	24.010	< 2e-16	***
ASIAN_PERC	129.8117	5.1790	25.065	< 2e-16	***
AA_PERC	99.3448	4.1430	23.979	< 2e-16	***
HISPANIC_PERC	115.8071	4.0426	28.646	< 2e-16	***
MULTIRACE_PERC	96.4154	7.1710	13.445	< 2e-16	***
CAUCASIAN_PERC	115.7613	4.0393	28.658	< 2e-16	***
DISABILITY_PERC	-50.7046	1.5058	-33.673	< 2e-16	***
ENGLSIH_PRO_PERC	-3.1062	2.5963	-1.196	0.231562	
PLI	-19.7937	0.9103	-21.744	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.65 on 11217 degrees of freedom
Multiple R-squared: 0.3839, Adjusted R-squared: 0.3808
F-statistic: 124.8 on 56 and 11217 DF, p-value: < 2.2e-16

In our third attempt to explain variation in graduation rates, our efforts were much more fruitful with a new R-squared value of 38.12%. We decided to engineer several variables from our graduation dataset in order to try and extract more information. For example, we decided to take a look at the demographic characteristics of the school districts. For example, from above, our DISABILITY_PERC variable was one that was derived from our original dataset. It was calculated by taking the total number of disabled students within a school district and dividing by the total number of students in that district. This gave us the fraction of disabled students per district. As we can see, not only was the parameter statistically significant, but it also had a negative effect on our graduation rate.

Though our model accounted for a significant proportion of variation in graduation rates, we believe that this model can be improved upon. Future investigation could be conducted to determine other variables that contribute to high school graduation rates. These analyses may involve merging district level graduation rate data files with other publically available data. Prospective researchers are encouraged to improve upon our baseline model and graph their data over geographic space or use other methods of data visualization.

***It is important to note that these benchmarks are simply starting points and we expect them to be improved by adding more variables to your model.**

Data Source: U.S. Department of Education. Provisional Data File of SY2011-12 District Level Four-Year Regulatory Adjusted Cohort Graduation Rates (ACGR). Retrieved from https://inventory.data.gov/dataset/73021f54-f5bf-4f34-86dd-5497ba4bdcfc/resource/32b21f42-b80b-4b1b-ad50-b361c5cecd99

** Minor data preparation was performed. Some of the graduation rates were reported with ranges (ie. 80-84) or with characters in front of them (ie. G82). For those with ranges, we took the average of the range, and for those with characters, we parsed them out to leave the value completely numerical.

Additional Data that may be merged with District Level ACGR data file(s)

- (i.e., using the LEAID unique identifier variable):

Census Data

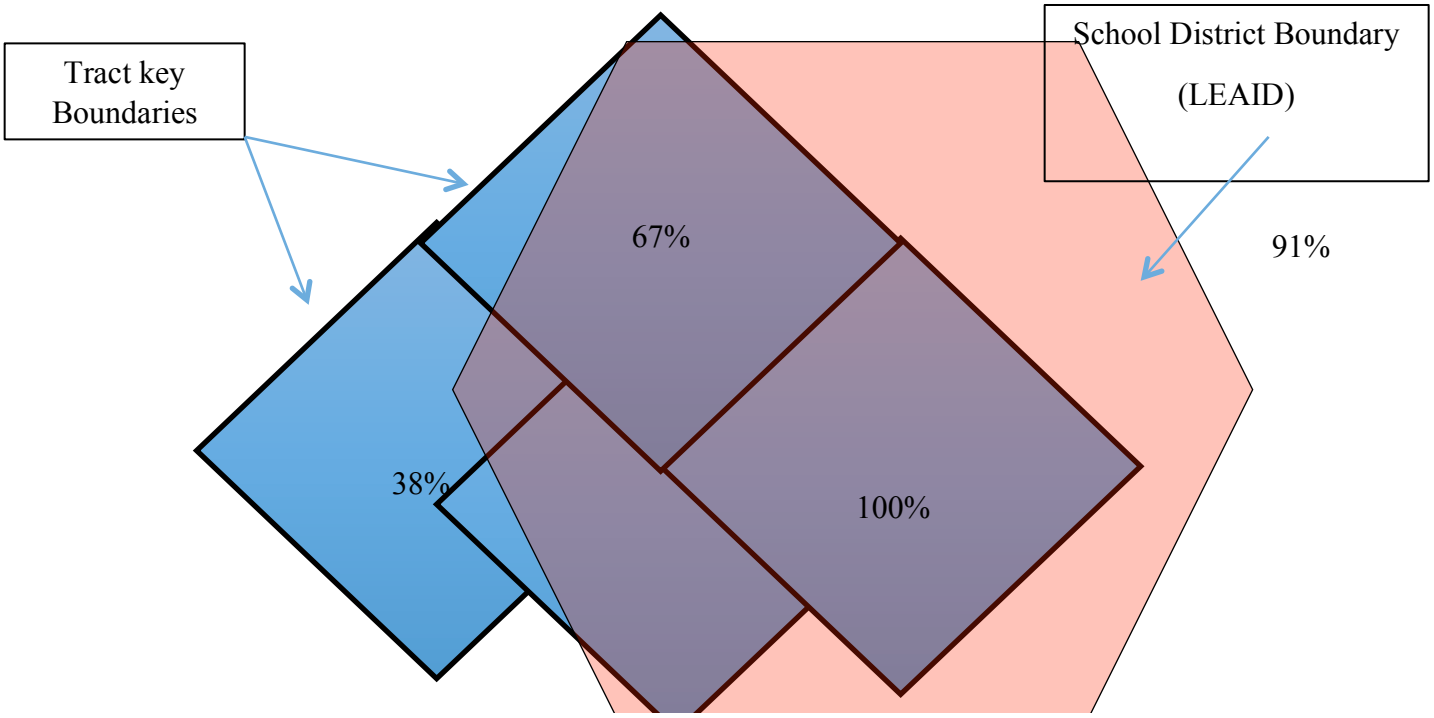
We are also providing a large file filled with CENSUS data for almost 90% of the school districts found in our graduation dataset mentioned previously.

How did we gather this census data for each school district?

Census data is collected and recorded on a 'TRACT' level. According to the census site, '**Tracts** are small, relatively permanent statistical subdivisions of a county or equivalent entity that are updated by local participants prior to each decennial census as part of the Census Bureau's Participant Statistical Areas Program. ' - https://www.census.gov/geo/reference/gtc/gtc_ct.html

Using geographic information for school districts AND tract boundaries, we were able to discover *which* tracts overlapped into certain school districts.

Example: For the sake of explanation, lets assume the blue squares are Census Tract boundaries, and the red hexagon is a school district boundary. **We have calculated the overlap percentage for every tract key (State ID + County ID + Tract Code) that overlaps any part of the school district boundary.**



The file *SD_TRACT_MAPPING2010.csv* will contain the overlap percentages for every LEAID (school district). In order to map this back to the CENSUS file for 2010, the join key will be the concatenation of (State,County,Tract Code). **For our POC, we chose the corresponding CENSUS data for the tract that had the highest percentage overlap**. To have better and more accurate resolution, however, we encourage you to try and account for each tract that overlaps, as they contribute additional information to variations found within an LEAID. For example, if one were looking at median household income for a particular school district, it *might* be advantageous to do a weighted average of the over lapping income levels to get a better idea of the demographics surrounding that district.

Income Example;

School District	Tract	Overlap Percentage	Median HH Income
A	10001	41%	$45,000
A	10002	36%	$56,000
A	10003	17%	$41,000

Taking into account the weighting and overlap, our newly calculated Median HH Income for School District 'A' would be (.41*45000 + .36*56000 + .17*41000 = $45,580)

Below are mapping keys you may find useful when using the datasets.