Similar Schools Comparison Tool

Gower Method Documentation

What is the Gower Method?

The Gower Method is a metric that is able to measure the similarity or dissimilarity of two objects across multiple characteristics (Gower, 1971). It is able to accommodate both continuous and categorical data and produces a single coefficient between 0 and 1. This makes it an ideal measure for calculating a “match” percentage between schools because it can account for school characteristics that can be described numerically (e.g., percentage of students that are economically disadvantaged) and characteristics that are classifications (e.g., urbanicity).

How does the Gower Method work?

The Gower coefficient is the weighted average dissimilarity of all characteristics considered. Two objects that are completely identical have a coefficient of 0 (no dissimilarity), whereas two objects that are as dissimilar as possible have a coefficient of 1. Subtracting the coefficient from 1 converts it into a similarity score.

The way dissimilarity is calculated for each variable is dependent on the type of data (i.e., numerical or categorical). For numerical data:

- The coefficient is calculated by taking the absolute value of the difference between two characteristics and divide by the total range.
- For example, suppose the economically disadvantaged rate for School A and B were 45% and 53% respectively. Additionally, across all schools, the minimum rate was 5% and the maximum rate was 100%.
    - The coefficient for this characteristic would be (.53 - .45) / (1 - .05) = .08

For categorical data:

- 0 if the values are the same, 1 if they are different
- For binary variables are one value is rare, the coefficient is 0 only if the values are the same *and* equal the rare condition (e.g., only a handful out of 1800 schools have migrant students).

OSDE uses the daisy function in the “cluster” R package to calculate Gower coefficients (Maechler et al., 2022). Each accountability model type (i.e., PK – 3 schools, elementary schools, middle schools, and high schools) is run separately so that only schools of like type are compared. The output is saved as a .csv file where each row is a school, the school it is being compared to, a similarity coefficient between 0 and 1, and a square root transformation of the similarity coefficient. The square root transformation was included because the minimum observed coefficients were generally between .4 and 5., even amongst the schools that would intuitively seem to be the most

dissimilar. The square root transformation has the effect of stretching out the lower end of the distribution so that the observed range of similarity scores aligns more closely with a 0 to 1 scale.

For specific technical documentation on the daisy function in R or the cluster package in general, please see https://CRAN.R-project.org/package=cluster.

How were the variables and their weights determined?

At the time of the development of this tool, the most recent report card was from the 2018 – 2019 school year. Therefore, all data used to determine which variables to include and their respective weights were from this year. All analyses were done in R.

The primary goal of the Similar Schools Comparison Tool is to provide valuable information to education stakeholder by allowing the comparison of schools similar in demographic characteristics on their accountability indicators. If one school is outperforming other schools with similar characteristics, the other schools may look to see what that school is doing that may be causing the improved performance (e.g., what curriculum are they using)? Likewise, if a school is underperforming compared to similar schools, that may be a signal that different strategies may be needed at that school. Thus, the set of characteristics were chosen based on those that might be the most correlated with student performance. This set was as follows:

- Percentage by sex
- Percentage by race/ethnicity
- Percentage of EL students
- Percentage of student with disabilities
- Percentage of students identified as gifted and talented
- Percentage of students not enrolled for a full academic year (NFAY)
- Percentage of economically disadvantaged students
- The average number of students enrolled per grade level
- The number of grade levels served
- Whether or not a school serves free lunch to all students
- Whether or not a school has any students identified as immigrants
- Whether or not a school has any students identified as migrants
- Whether or not a school has any students placed there by virtue of court order or otherwise in a full-time residential facility
- The Rural Urban Commuting Area Code as determined by the USDA (https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/)
- Urbanicity status as determined by the US census (i.e., City, Suburb, Town, or Rural)
- The immigrants, migrants, and placement via court order variables were binary yes/no variables instead of percentages because the majority of schools had none, making the presence of any a more meaningful method of differentiation.

Other variables were initially considered (e.g., bilingual status), but were discarded due to either being so highly correlated with one or more of the above variables that they were basically redundant. Other variables were not included because they were considered to be a function of the organizational structure of the school as opposed to a characteristic of the student population (e.g., charter school status).

A logistic ordinal regression was performed to predict a school's classification of their overall grade on these variables. Regression coefficients were converted into proportional odds ratios to assess the importance each variable had in predicting a school's letter grade (Appendix A). For example, the odds ratio of percentage of NFAY students was 2.00. This means that for every standard deviation increase in NFAY percentage (approximately 10%), the likelihood of a school receiving a B versus and A (or a C versus a B, and so on) doubles. The regression was run a second time reversing the order of the letter grade variable so that the odds ratio represented the likelihood of moving up a letter grade as opposed to down. This was done so that the magnitude of impact could be consistently converted to weights regardless of whether the impact was positive or negative.

The final set of odds ratios rounded to the nearest tenth become the weights used in the Gower method.

References

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857-871.

Maechler M., Rousseeuw P., Struyf A., Hubert M., Hornik K. (2022). *cluster: Cluster Analysis Basics and Extensions.* R package version 2.1.4

Appendix A

Table 1. Odds ratios of school characteristics with 95% confidence intervals (higher ratio means increased odds of a **lower** grade).

	Odds Ratio	CI Low	CI High
EL Percent	**1.8**	1.4	2.2
Hispanic/Latino Percent	0.9	0.7	1.1
Male Percent	**1.2**	1.1	1.3
Black Percent	**1.6**	1.3	1.8
Two Or More Races Percent	**1.0**	0.9	1.1
American Indian Percent	**1.2**	1.0	1.4
Asian/Pacific Islander Percent	0.8	0.7	0.9
Students with Disabilities Percent	**1.0**	0.9	1.1
Gifted and Talented Percent	0.9	0.8	1.1
NFAY Percent	**2.0**	1.7	2.4
Economic Disadvantage Percent	**1.7**	1.4	2.0
All students receive lunch	**1.5**	1.1	1.9
Enrollment Per Grade	0.3	0.1	1.4
Number of Grades Served	0.8	0.8	0.8
Has immigrants	0.7	0.5	0.9
Has placed students	0.7	0.5	0.9
Has migrants	0.5	0.3	1.0
Rural Urban Commuting Area Code	**1.0**	1.0	1.0
Town	**1.7**	1.1	2.6
Suburb	**1.7**	1.1	2.6
Rural	**1.8**	1.2	2.9

**Bold numbers indicate the final weights used in the Gower Method.*

Table 2. Odds ratios of school characteristics with 95% confidence intervals – Run 2 (higher ratio means increased odds of a **higher** grade).

	Odds Ratio	CI Low	CI High
EL Percent	0.6	0.4	0.7
Hispanic/Latino Percent	**1.1**	0.9	1.5
Male Percent	0.8	0.7	0.9
Black Percent	0.6	0.5	0.7
Two Or More Races Percent	**1.0**	0.9	1.1
American Indian Percent	0.8	0.7	1.0
Asian/Pacific Islander Percent	**1.3**	1.2	1.4
Students with Disabilities Percent	**1.0**	0.9	1.1
Gifted and Talented Percent	**1.1**	0.9	1.2
NFAY Percent	0.5	0.4	0.6
Economic Disadvantage Percent	0.6	0.5	0.7
All students receive lunch	0.7	0.5	0.9
Enrollment Per Grade	**3.1**	0.7	14.5
Number of Grades Served	**1.2**	1.2	1.3
Has immigrants	**1.5**	1.1	2.0
Has placed students	**1.5**	1.1	2.0
Has migrants	**1.9**	1.0	3.5
Rural Urban Commuting Area Code	**1.0**	1.0	1.0
Town	0.6	0.4	0.9
Suburb	0.6	0.4	0.9
Rural	0.5	0.3	0.9

**Bold numbers indicate the final weights used in the Gower Method.*

The percentage of female students, the percentage of white students, and City urbanicity status are not present in the model because they are already implicitly represented by the other variables (e.g, if a school is not in a Town, Rural area, or Suburb, it must be in a City).