Methodology to Compare Districts and Schools: A Technical Report

January 18, 2019

Prepared by

Office of Data, Research and Evaluation
Nebraska Department of Education

Table of Contents

Note of Caution

The methodology described in this document represents *one* approach to constructing a group of similar peers for each school and school district in Nebraska. Other methods could also be used. As such, we caution readers to interpret the similar peer information with care. When evaluating school and school district data, persons should consider a mix of reference points as a means of triangulation. Other reference points might include, for example: the state average, statistics for those schools and school districts geographically closest, statistics for schools receiving similar supports and services, and those with the most similar membership counts.

Limitations

Developing similar peer groupings is designed to enable users to conduct more thoughtful comparative analysis. Despite the benefits to this approach, there are limitations to the use of any grouping methodology. Specific limitations to the approach employed here include:

- The similar peer calculation does not include a measure of geographic distance (although users can select geographic distance as a separate parameter using the NEP compare feature). Many schools and school districts tend to compare themselves with surrounding schools and school districts. The similar peer method does not necessarily include geographically close districts in the comparison grouping because neighboring districts might not truly be the "most similar" districts in the state. On the other hand, some variables included in the similar peer calculation tend to reflect regional conditions.
- The similar peer method deliberately selects only the 12 schools or school districts "most similar" as the standard for comparison. However, some schools and districts are more "unique" than others. In some cases, "similarity" to other schools or school districts – even among peers – can be large.
- It is also true that some schools or school districts tend to look like many other schools or school districts, so the cutoff of 12 captures those schools or school districts that are extremely similar according to the chosen dimensions. Still, schools or school districts can closely resemble many other schools or school districts beyond the cutoff of 12.

Acknowledgments

- Dr. Ashok Samal, Professor, Department of Computer Science and Engineering, University of Nebraska-Lincoln (UNL) for providing invaluable guidance on variable identification, data management, and clustering approaches.
- David Drozd, Research Coordinator, Center for Public Affairs Research (CPAR), University of Nebraska at Omaha (UNO) for providing expert insight into Census data and procuring the desired Census data for this project.
- Office of Policy and Research, Ohio Department of Education, for sharing their experience and method of determining similar districts in their state.
- Subject Matter Experts at the Nebraska Department of Education (NDE) for sharing their knowledge on the appropriateness of specific data elements for this project.

Keywords: *Similar Districts; Similar Schools; Nebraska Education Profile; Census Data; Euclidean Distance; Geographic (Haversine) Distance*

Introduction

The Nebraska Education Profile (NEP) website has been undergoing major enhancements, and thus the need to identify and compare similar peer districts and schools. This would provide utility for any given district or school as they evaluate their performance relative to that of the entire state, and relative to that of other districts or schools that are similar to them on a variety of measures – peers. Additionally, groups of districts or schools that are geographically close to each other are also determined to allow for comparisons between districts or schools within the same geographical area. This technical report details the methodology behind these similar peers and geographic groupings.

Similar Peer Districts and Schools

Design and Methods

In order to operationalize "similarity," a combination of variables that uniquely describes each district or school was identified. These variables were selected due to their relevance, availability, and persistence. Table 1 describes the list of 27 variables that were selected to describe any given district or school.

Table 1. Variables used to compare similarity between districts and schools.

Variable	Description	Source
Membership	Number of students enrolled	NDE
Attendance Rate	Average student attendance rate	NDE
Graduation Rate	4-year graduation rate for the 2016-2017 cohort	NDE
FRL Rate	Percentage of free-and-reduced lunch students	NDE
Minority Rate	Percentage of non-White students	NDE
Homeless Rate	Percentage of homeless students	NDE
LEP Rate	Percentage of English language learners	NDE
Migrant Rate	Percentage of migrant students	NDE
ELA Percent Proficient	Percentage of students proficient in ELA	NDE
Math Percent Proficient	Percentage of students proficient in Math	NDE
Science Percent Proficient	Percentage of students proficient in Science	NDE
Teachers With Masters Percent	Percentage of teachers with at least a Master's degree	NDE
Average Years Teaching Experience	Average number of years taught by teachers	NDE
Unduplicated Suspensions	Number of students with suspensions	NDE
Unduplicated Expulsions	Number of students with expulsions	NDE
Land Valuation	Annual land valuation sent out from the County Treasurer's office of the district	NDE

Variable	Description	Source
Per Pupil Cost by Average Daily Membership	Total annual costs divided by the average daily membership for the district	NDE
Grand Total of All Receipts	Amount of all receipts/revenue received by the district in a school year	NDE
Median Household Income	Median household income in the past 12 months (in 2016 inflation-adjusted dollars)	Census-ACS 2012-2016
Per Capita Income	Per capita income in the past 12 months (in 2016 inflation-adjusted dollars)	Census-ACS 2012-2016
Gini Index	Gini index of income inequality	Census-ACS 2012-2016
Percent Age 25+ With Bachelor's Degree or More	Percent of population 25 years and over with at least a Bachelor’s degree	Census-ACS 2012-2016
Labor Force Participation Rate	Percent of population 16 years and over in the labor force	Census-ACS 2012-2016
Unemployment Rate	Percent of population 16 years and over who are unemployed	Census-ACS 2012-2016
Total Population	Population in the district	Census 2010
Land Area	Area in square miles	Census 2010
Population Density	Density per square mile of land area	Census 2010

In creating the district and school data sets from various data sources, a number of challenges surfaced. First, the latest data from NDE was the 2016-2017 school year, while the latest data from the Census was from 2010, and from 2012-2016. Although the Census data lagged behind NDE’s data on the districts and schools, the Census data was still used since the variables described community characteristics (e.g., median household income, land area, etc.) that would likely not have changed as frequently as the school characteristics (e.g., membership, attendance rate, etc.).

Second, the Census data was only collected at the district-level, and not at the school-level. However, since the community characteristics of a given district would reflect that of the schools within the district, the same Census data was used at the school-level. This implied that all schools within the same district would, for example, have the same unemployment rate as that of the district. Three pieces of finance data were also collected at the district-level only by NDE: land valuation, per pupil cost by average daily membership, and grand total of all receipts. By the same logic aforementioned, district-level information was used for the schools within the same district.

Third, there were a number of districts that were consolidated after the Census data was collected. In these cases, the originating districts were first identified in the Census data, and the average values of the Census variables were then calculated to inform the Census variables for the new consolidated district.

Once the aforementioned decisions were made, a data split was performed on only the school data file. The school data file was split into three separate data files to reflect the differences among elementary, middle, and high schools. The number of students with expulsions was found to have very little variability across the schools (due to many zero values) and was thus removed from all school data files. Only one variable was not available to describe the elementary and middle schools, namely, graduation rate which was only applicable to high school students. With three school data files, and one district data file, the analyses to identify similar districts and schools commenced.

Analytic Approach

Each district or school was compared to every other district or school by using a distance measure between each pair of districts or schools. This Euclidean distance measure was calculated as a summary index using the formula shown below:

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

In the formula above, *d* represents the distance between any two districts or two schools *x* and *y* on each variable *i* (i.e., every variable shown in Table 1). Due to the wide differences in the ranges of values across the variables, each variable was scaled prior to computing the Euclidean distance.

Thus, for each district or school, the districts or schools with the shortest distances to it are grouped together. This is because the shorter the Euclidean distance between two districts or two schools, the more similar they are.

Geographical Area

Design and Methods

The addresses for each district and school building were first converted into latitude and longitude information. Once this was done, the geographic distance between every pair of districts and every pair of schools was calculated using the Haversine distance measure. Note that the school data file was split into three separate data files to ensure that similar school types were being compared to each other. For example, elementary schools were only compared with other elementary schools in terms of geographic distance. The same held true for middle schools and high schools as well.

Table 2. Variables used to describe geographic location for districts and schools.

Variable	**Description**	**Source**
Latitude	North-South geographic coordinate	Google Maps
Longitude	East-West geographic coordinate	Google Maps

Analytic Approach

Each district or school was compared to every other district or school by using a geographic distance measure between each pair of districts or schools. This Haversine distance represents the distance between two coordinates on a sphere and was calculated using the formula shown below:

$$d_{hav}(x,y) = 2r\ \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi_y - \varphi_x}{2}\right) + \cos(\varphi_x)\cos(\varphi_y)\sin^2\left(\frac{\lambda_y - \lambda_x}{2}\right)}\right)$$

In the formula above, *d* represents the geographic distance between any two districts or two schools *x* and *y*, with φ representing the latitude and λ representing the longitude.

Results

The results of this work can be found as an interactive display in the Nebraska Education Profile website: http://nep.education.ne.gov/. Once a district or school is selected from the dropdown menu on the main page, the "Compare" feature can then be selected to show 10 other districts or schools that are most similar or geographically closest to the referent district or school. For questions or comments regarding the use of this feature, please reach out to NDE.Research@nebraska.gov.

Contributors

This research effort was conducted by the following researchers at the Office of Data, Research and Evaluation at the Nebraska Department of Education:

- Matt Hastings, Ph.D., Senior Administrator
- Hongwook Suh, Ph.D., Psychometrician Lead
- Justine Yeo, Statistical Research Analyst
- Kunal Dash, Statistical Research Analyst
- Fisayo Adeniyan, Research Assistant

Appendix

All distance calculations were computed using R, a statistical software. The syntax is shown in the tables below. While only the syntax for the district data is presented, the same syntax was also applied to all school data files.

Table 3. Syntax for calculating Euclidean distances for every pair of district.

```
###Euclidean Distance
###District Data

#install.packages("ggplot2")
library(ggplot2)
#install.packages("factoextra")
library(factoextra)
#install.packages("xlsx")
library(xlsx)



getwd()
setwd("District Data")
getwd()
district <- read.csv("District Data v0.09.csv")
head(district)
#district <- na.omit(district)

district[,-c(1)] <- scale(district[, -c(1)])
head(district)

districtdistance <- dist(district, method="euclidean")
as.matrix(districtdistance)
as.matrix(districtdistance)[1:6, 1:6]
distanceframe <- round(as.matrix(districtdistance), 5)
str(distanceframe)

fviz_dist(districtdistance)

write.csv(distanceframe, "District Euclidean Distance.csv")
```

Table 4. Syntax for converting addresses to latitude and longitude coordinates, and for calculating Haversine distances for every pair of district.

```
###Geocoding
###District Addresses Data

#Install necessary packages
#install.packages("tidyverse")
```

```
library(tidyverse)
#install.packages("ggmap")
library(ggmap)
#install.packages("geosphere")
library(geosphere)
#install.packages("ggplot2")
library(ggplot2)
#install.packages("xlsx")
library(xlsx)

#Set working directory
getwd()
setwd("Geographic Distance")
getwd()

#Import data with addresses
adddistrict <- read.csv("District Address v0.01.csv", stringsAsFactors = FALSE)
head(adddistrict)
adddistrict <- na.omit(adddistrict)

#Convert addresses to longitude and latitude
?mutate_geocode
geodistrict <- mutate_geocode(adddistrict, Location)
head(geodistrict)

#Check status of query counts from Google Maps (limited to 2500 queries per day)
geocodeQueryCheck()

#Export data with longitude and latitude columns appended
write.csv(geodistrict, "District Geocode v0.01.csv")

#Import data with longitude and latitude columns only
district <- read.csv("District Geocode for Distances v0.01.csv")
head(district)

#Drop agency name which is the first column in the data
district2 <- district[,-c(1)]
head(district2)

#Calculate distance between every pair
distance <- distm(district2, fun=distHaversine)

#Convert distances into a matrix
as.matrix(distance)
as.matrix(distance)[1:6, 1:6]
str(distance)

#Export matrix of distances
write.csv(distance, "District Geographic Distance v0.01.csv")
```