Geographical Database

Mahmoud Hamsho, Winnie Lee, Tanya Piplani, Jeffery Yu

Introduction

Inspired by the primary election data of California state[1], we are interested in exploring the geographical dataset and finding any insights between landscape and other topics, like political orientation, income, health, education, or races in the different neighborhoods. While searching for the applicable data source, we found the practical data attribute, Census Tract, among several datasets. According to the Census Bureau's Participant Statistical Areas Program, “Census Tracts are small, relatively permanent statistical subdivisions of a county or equivalent entity that are updated by local participants prior to each decennial census[2].”

In addition, Census Tract is a suitable geographical size for representing the neighborhood which meet our goal for this project. It’s not too detailed like the specific longitude and latitude and not too board like the area at the county or city level. Furthermore, compared to the ZIP code, using Census Tract could divide the neighborhood into the scale of streeblock. To be more specific, the population under zip code 94704 is about 22,000 while covering 10 different Tract areas which population are ranged from 28 to 6800[3]. Therefore, Census Tract could be the good candidate as primary key for connecting the multiple datasets.

Data Source

We will be taking the Primary election precinct data for California from the Statewide Database[4]. The datasets are in the form of of csv files. Available Precinct data files are SOV, REG, ABS, MAIL, POLLV and VOTE and are available by different Precinct type (rgprec, rrprec, srprec, ssprec, svprec, mprec etc.) We will be connecting the above mentioned census data with other entities like Income, Race, demographics, education using the tract field.

1. Income, race and gender
    a. Household Income and Incarceration for Children from Low-Income Households by Census Tract, Race, and Gender [5]. This table reports predicted outcomes for children by Census tract, race and gender. Each Census tract is uniquely identified by three identifiers – state, county, and tract (2010 FIPS codes). The

[1] http://statewidedatabase.org/d10/p16.html
[2] https://www.census.gov/geo/reference/gtc/gtc_ct.html
[3] http://proximityone.com/ziptractequiv.htm
[4] http://statewidedatabase.org/d10/p16.html
[5] https://opportunityinsights.org/wp-content/uploads/2018/10/Codebook-for-Table-1-1.pdf

data is organized long on Census tract and wide on race and gender, so that there is exactly one row per tract. It provides data for children born between 1978 and 1983.

b. All Outcomes by Census Tract, Race, Gender and Parental Income Percentile[6] - This table reports predicted outcomes for children by Census tract, race and gender. Each Census tract is uniquely identified by three identifiers – state, county, and tract (2010 FIPS codes). The data is organized long on Census tract and wide on race, gender and parent income percentile rank, so that there is exactly one row per tract. It is similar to the above dataset but includes more types of outcomes.

2. Life Expectancy and Income
    a. County-level life expectancy estimates for men and women, by income quartile[7] - This table reports life expectancy point estimates and standard errors for men and women at age 40 for each quartile of the national income distribution by county of residence. Both race adjusted and unadjusted estimates are reported. Estimates are reported for counties with populations larger than 25,000 only.

3. Neighbourhood Characteristics
    a. Neighborhood Characteristics by Census Tract[8] - This table provides tract-level covariates used throughout the paper or shown in the Opportunity Atlas as neighborhood characteristics. Each Census tract is uniquely identified by three identifiers – state, county, and tract (2010 FIPS). These covariates are constructed based on publicly available sources. It consists of the statistics for each track, such as percentage of foreign born residents, employment rate etc.

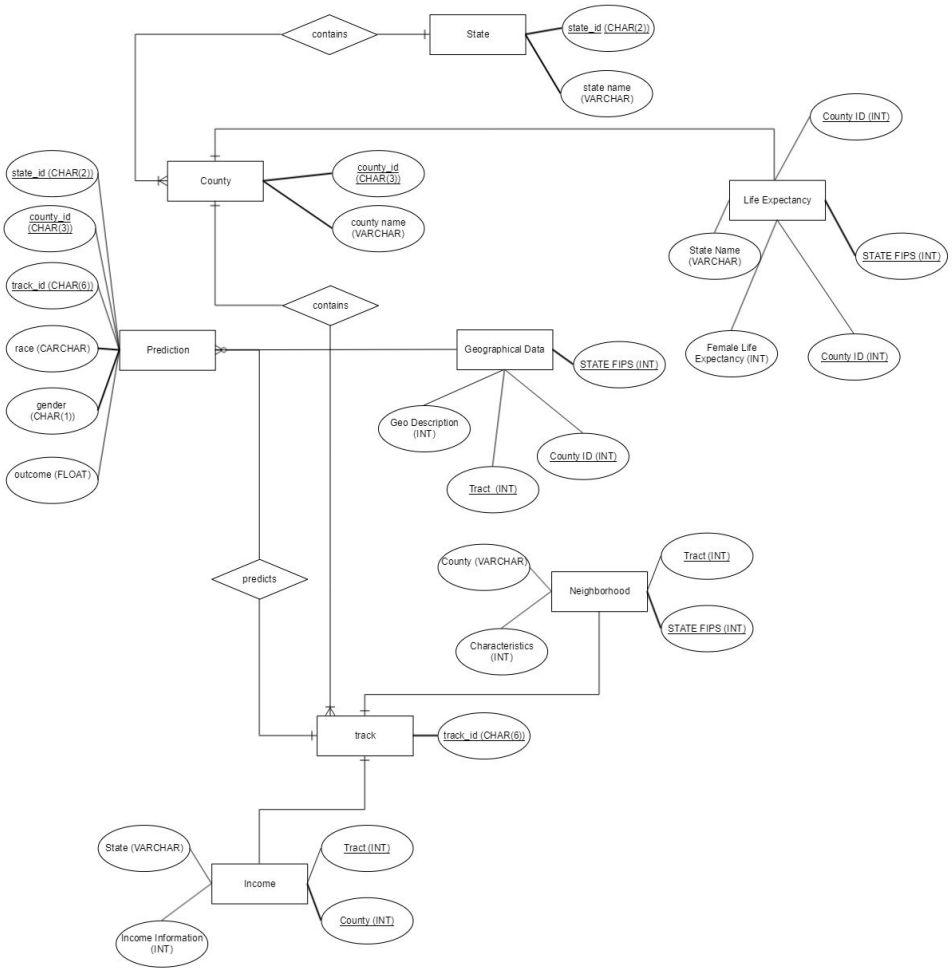
4. Education
    a. Data files from a variety of California Department of Education data collections[9] are available to educators, researchers, and the public. Data files can be used to compare educational data with other data sets. This 14-digit code is the official, unique identification of a school within California. The first two digits identify the county, the next five digits identify the school district, and the last seven digits identify the school. Please note that a CDS code ending in '0000000' indicates a district record not a school.

5. Geography
    a. Geographical dataset are collected from the United States Census Bureau that contain entity codes that can be linked to the other datasets via tract and state

[6] https://opportunityinsights.org/wp-content/uploads/2018/10/Codebook-for-Table-4.pdf
[7] https://opportunityinsights.org/wp-content/uploads/2018/04/health_ineq_online_table_11_readme.pdf
[8] https://opportunityinsights.org/wp-content/uploads/2018/10/Codebook-for-Table-9.pdf
[9] https://www.cde.ca.gov/ds/dd/
[10] https://www.census.gov/geo/maps-data/data/tiger-line.html

codes. The files are made up of DBF and Shapefiles that will be used to map and visualize the other datasets. The spatial data is digitized to the TIGER system[10].

ERD (Entity-Relationship Diagram)



Data Dictionary:

Column	Data Type	Description
Income		
State	Varchar	A string representation of a specific state name
Income Information	Integer	Household income data

Tract	Integer	Geographical boundary lines
County	Integer	Numerical representation of county

Column	**Data Type**	**Description**
Neighborhood		
State FIPS	Integer	A five-digit Federal Information Processing Standards code
Characteristics	Integer	Statistical description of neighborhood demographics
County	Varchar	Name of county

Column	**Data Type**	**Description**
Life Expectancy		
State Name	Varchar	Name of state
State FIPS	Integer	A five-digit Federal Information Processing Standards code
County ID	Integer	Numerical representation of county
Female Life Expectancy	Integer	Statistical measure of life expectancy

Column	**Data Type**	**Description**
State		
State ID	char(2)	Two-digit state 2010 FIPS code
State name	varchar	Name of state

Column	Data Type	Description
County		
County ID	Char(3)	Three-digit county 2010 FIPS code
County name	Varchar	Name of county

Column	Data Type	Description
Track		
State ID	char(2)	Two-digit state 2010 FIPS code
County ID	Char(3)	Three-digit county 2010 FIPS code
Track ID	Char(6)	Six-digit tract 2010 FIPS code. A census tract code may not be used more than once in a single county, but it may be used again in a different county in the same state or in a county in a different state. Therefore, a particular census tract within the nation must be identified by: its state, its county, and its tract code.

Column	Data Type	Description
Prediction		
State ID	Char(2)	Two-digit state 2010 FIPS code
County ID	Char(3)	Three-digit county 2010 FIPS code
Track ID	Char(6)	Six-digit tract 2010 FIPS code
Race	Varchar	white, Black, Hispanic, Asian, Native American(natam)

Gender	Char(1)	0=Male, 1=Female
Household Income Rank	Float	Mean percentile rank (relative to other children born in the same year) in the national distribution of household income (i.e. own earnings and spouse's earnings) measured as mean earnings in 2014-2015 for the baseline sample

Reference

- Statewide database, http://statewidedatabase.org
- The Opportunity Altas, https://www.opportunityatlas.org