Phuong’s Working Document

Table of Contents

# Description of the document

This is the Opportunity Mapping 2.0 Technical Document produced by the Phuong Tseng. The intention is to capture changes and developments in the 2019 version.

# The Methodology Document and Spreadsheet

[2019 Opportunity Mapping Indicators and Measures](https://docs.google.com/spreadsheets/d/1zlJliXjgMCKbtgz4-mM4lPTMmCqCjBJITkYPG7aIv0k/edit#gid=0).

[OM\_methodology\_v4\_Nov30.pdf]

## Set-up

## A. The Domains

In 2019, there are 5 domains: education, economic & mobility, housing and neighborhood, conduit, and social capital. The social capital domain is a new domain in 2019.

### 1. Education Opportunity Indicators

This year, the education domain added a new indicator called Early Childhood Participation Rate or Pre-K. Another indicator, adult with bachelor’s degree was moved from the education domain to the economic & mobility domain in 2019.

edu\_list <-  
 c(  
 "fips",  
 "CountyID.x",  
 "TOTPOP.x",  
 "math\_prof",  
 "read\_prof",  
 "grad\_rate",  
 "pct\_not\_frpm",  
 "z\_math\_prof",  
 "z\_read\_prof",  
 "z\_grad\_rate",  
 "az\_pct\_not\_frpm",  
 "county\_name.x",  
 "HD01\_VD04",  
 "HD01\_VD03",  
 "ratio",  
 "ratio2",  
 "z\_preK"  
 )

### 2. Economic & Mobility Opportunity Indicators

There are a few changes to this domain in 2019. The adult with bachelor’s degree was added to this domain, median household income, and median household value. Other indicators such as the commuting time and entry-level jobs’ measures were changed to TCAC’s measures. A new indicator, school district revenue per capita, was added to capture the extent of municipal hoarding. Due to reliability issues of municipal data, school district boundary was used as a proxy instead.

econ\_list <- c(  
 "fips",  
 "CountyID.x",  
 "TOTPOP.x",  
 "total\_pop\_2017",  
 "below\_200\_pov\_2017.x",  
 "moe\_below\_200\_pov\_2017.x",  
 "pct\_below\_pov\_2017",  
 "moe\_pct\_below\_pov\_2017",  
 "pct\_below\_200\_pov\_2017.x",  
 "pct\_assist\_2017",  
 "med\_hhincome\_2017" ,  
 "moe\_med\_hhincome\_2017" ,  
 "employed\_pop\_20to60\_2017",  
 "pct\_employed\_20to60\_2017",  
 "home\_value\_2017" ,  
 "moe\_home\_value\_2017",  
 "pct\_bachelors\_plus\_2017",  
 "above\_200\_pov\_2017",  
 "pct\_above\_200\_pov\_2017",  
 "tot\_hh\_2017",  
 "moe\_tot\_hh\_2017",  
 "moe\_pct\_long\_commute\_2017",  
 "moe\_assist\_2017",  
 "moe\_long\_commute\_pct",  
 "long\_commute\_pct",  
 "low\_wage\_med\_distance" ,  
 "jobs\_lowed" ,  
 "rural\_flag",  
 "az\_pct\_assist\_2017" ,  
 "az\_pct\_employed\_20to60\_2017",  
 "z\_home\_value\_2017" ,  
 "z\_pct\_bachelors\_plus\_2017" ,  
 "az\_pct\_long\_commute\_2017",  
 "z\_jobs\_lowed" ,  
 "Econ\_Domain",  
 "z\_sdrevpcap",  
 "sdrev",  
 "sdrevpcap",  
 "sd\_totpop"  
 )

### 3. Housing & Neighborhood Opportunity Indicators

The housing and neighborhood opportunity domain has two new environmental indicators pulled from CalEnviroScreen (i.e. pm25, lead).

housing\_list <-  
 c(  
 "fips",  
 "CountyID.x" ,  
 "TOTPOP.x",  
 "county\_name.x",  
 "below\_200\_pov\_2017.y",  
 "moe\_below\_200\_pov\_2017.y",  
 "pct\_below\_200\_pov\_2017.y",  
 "pm25",  
 "pct\_pm25",  
 "toxRelease",  
 "pct\_toxRelease",  
 "lead\_pctl",  
 "pct\_lead\_pctl" ,  
 "Grocery",  
 "z\_Grocery" ,  
 "az\_Grocery",  
 "P\_INSURED" ,  
 "az\_insurance" ,  
 "H\_Crime",  
 "pct\_parks",  
 "az\_pct\_below\_200\_pov\_2017",  
 "az\_pct\_below\_200\_pov\_20172",  
 "az\_pct\_pm25",  
 "az\_pct\_toxRelease",  
 "az\_pct\_lead\_pctl" ,  
 "Housing\_Env\_Domain",  
 "test\_azcrime" ,  
 "azhealthcare" ,  
 "zparks"  
 )

### 4. Conduit

The Conduit domain has two indicators: median broadband download speed and percentage of single-parent households.

conduit\_list <-  
 c(  
 "fips",  
 "CountyID.x",  
 "TOTPOP.x",  
 "pct\_singleparent\_hh\_2017.y",  
 "moe\_pct\_singleparent\_hh\_2017.y",  
 "az\_pct\_singleparent\_hh\_2017",  
 "TOTPOP.y",  
 "Median\_bb",  
 "z\_broadband",  
 "z\_broadband2",  
 "Conduit"  
 )

### 5. Social Capital

This is our newest domain, which has the average distance to a religious institution, registered voters voting rate, and average distance to club membership and etc.

socap\_list <-  
 c(  
 "fips",  
 "CountyID.x",  
 "TOTPOP.x",  
 "pct\_singleparent\_hh\_2017.y",  
 "moe\_pct\_singleparent\_hh\_2017.y",  
 "az\_pct\_singleparent\_hh\_2017",  
 "Clubs",  
 "AVGDIS\_REL",  
 "reg\_vote",  
 "SOCIAL\_CAP",  
 "z\_regvoter",  
 "zreligious",  
 "zclubs"  
 )

### 6. Compile All Indicators Function

### 7. Calculate Domains

## B. Index Calculation

data$index <- (data$housing\_domain + data$edu\_domain + data$econ\_domain + data$Socap\_domain + data$Conduit\_domain)/5  
  
#df <- data %>% select(fips,housing\_domain,edu\_domain,econ\_domain,Socap\_domain, Conduit\_domain, index) %>% filter(is.na(index))

## C. Filters

### 1. Filtering Single parent families >= 30%

returns 471 records with 8 NAs

#filter\_function <- function(data, variable1, variabl2, value, value2){  
# data$variable1[which(data$variable2)] <- value2  
# return(data$variable)  
#}  
  
data$SPF\_GT\_30[which(data$pct\_singleparent\_hh\_2017.y<0.3)] <- 0  
  
data$SPF\_GT\_30[which(data$pct\_singleparent\_hh\_2017.y>=0.3)] <- -1 #471 records  
  
data$flag\_spf <- ifelse(is.na(data$SPF\_GT\_30), 0, data$SPF\_GT\_30)  
  
summary(data$flag\_spf) #fixed NAs to 0, 471 records

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.0000 -1.0000 0.0000 -0.2983 0.0000 0.0000

### 2. Filtering Poverty (below 200 FPL) >= 30%

returns 418 records with 3 NAs

data$POVR200\_GT\_30[which(data$pct\_below\_200\_pov\_2017.x<0.3)] <- 0  
  
data$POVR200\_GT\_30[which(data$pct\_below\_200\_pov\_2017.x>=0.3)] <- -1 #418  
  
data$POVR200\_GT\_30 <- ifelse(is.na(data$POVR200\_GT\_30), 0, data$POVR200\_GT\_30)  
  
sum(data$POVR200\_GT\_30) #418

## [1] -418

### 3. Filtering Single parent >= 30% AND Poverty (below 200 FPL) >= 30%

data$SPF30\_P30[which(data$flag\_spf==0 | data$POVR200\_GT\_30==0)] <- 0  
  
data$SPF30\_P30[which(data$flag\_spf==-1 & data$POVR200\_GT\_30==-1)] <- -1  
  
summary(data$SPF30\_P30) #fixed NAs to 0

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.0000 0.0000 0.0000 -0.1849 0.0000 0.0000

sum(data$SPF30\_P30) #292

## [1] -292

### 4. High Divergence and population of Black and Latinx > 50%

load(here("data", "input\_DI.RData"))  
  
input\_DI$Flag\_HighDI\_Blk\_Lat[which(input\_DI$Black\_Latinx<=0.5 | input\_DI$divergence\_thresh<3)] <- 0  
  
input\_DI$Flag\_HighDI\_Blk\_Lat[which(input\_DI$Black\_Latinx>0.5 & input\_DI$divergence\_thresh==3)] <- -1  
  
data$fips <- paste0(0,data$fips)  
data <- merge(data,input\_DI, by="fips")  
sum(data$Flag\_HighDI\_Blk\_Lat) #201 no NAs

## [1] -201

### 5. High Divergence with population of Black and Latinx > 50% and poverty (below 200 FPL) >= 30%

data$DI\_Blk\_Lat\_POV30[which((data$Black\_Latinx < 0.5 & data$divergence\_thresh != 3) | (data$POVR200\_GT\_30 == 0))] <- 0  
sum(data$DI\_Blk\_Lat\_POV30)

## [1] NA

#filter\_na <- data %>% select(fips, county\_name,DI\_Blk\_Lat\_POV30,Black\_Latinx,divergence\_thresh,POVR200\_GT\_30) %>% filter(is.na(data$DI\_Blk\_Lat\_POV30))  
  
data$DI\_Blk\_Lat\_POV30[which((data$Black\_Latinx > 0.5 & data$divergence\_thresh == 3) & (data$POVR200\_GT\_30 == -1))] <- -1  
#320  
sum(data$POVR200\_GT\_30) #418 no NAs

## [1] -418

### 6. Final Filter

High Divergence with population of Black and Latinx > 50% and poverty (below 200 FPL) >= 30% OR Poverty (below 200 FPL) >= 30% and Single-parent family >= 30%

data$DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30[which((  
 data$Flag\_HighDI\_Blk\_Lat == 0 &  
 data$POVR200\_GT\_30 == 0  
 ) |  
 (  
 data$POVR200\_GT\_30 == 0 &  
 data$pct\_singleparent\_hh\_2017.y < 0.3  
 )  
 )] <- 0  
  
data$DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30[which((  
 data$Flag\_HighDI\_Blk\_Lat == -1 &  
 data$POVR200\_GT\_30 == -1  
 ) |  
 (  
 data$POVR200\_GT\_30 == -1 &  
 data$pct\_singleparent\_hh\_2017.y >= 0.3  
 )  
 )] <- -1  
  
data$DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30 <-  
 ifelse(  
 is.na(data$DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30),  
 0,  
 data$DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30  
 ) #fixed NAs  
  
sum(data$DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30) #320 records

## [1] -320

# Check filters

# Filter Function

#library(dplyr)  
categorize <-  
 data %>% dplyr::select(  
 "fips",  
 "index",  
 "DI\_Blk\_Lat\_POV30",  
 "SPF30\_P30",  
 "DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30"  
 ) %>% filter(DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30 == -1) %>%   
 mutate(category =  
 "Lowest Opportunity") #returns 320  
  
remaining <-  
 data %>% dplyr::select(  
 "fips",  
 "index",  
 "DI\_Blk\_Lat\_POV30",  
 "SPF30\_P30",  
 "DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30"  
 ) %>% filter(  
 DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30 == 0 &  
 (  
 fips != "06081984300" & fips != "06095253000" &  
 fips != "06095980000"  
 )  
 )  
  
remaining$DI\_Blk\_Lat\_POV30 <- 0  
  
nan\_records <-  
 data %>% dplyr::select(  
 "fips",  
 "index",  
 "DI\_Blk\_Lat\_POV30",  
 "SPF30\_P30",  
 "DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30"  
 ) %>% filter(fips == "06081984300" |  
 fips == "06095253000" |  
 fips == "06095980000")  
  
nan\_records$category\_2 <- NA

## D. Categorization with Filters

## E. Categorization without Filters

### 1. Graphs or Charts

Plotting index by opportunity categories but not sure why ggplot is not available

## F. Missing Values

These are records with NAs or missing values  
1. fips 06081984300 has NaN in pct\_pov\_below\_200 and pct\_singleparent\_hh  
2. fips 06081984300 (Mod) changed to NAs  
3. fips 06095253000 has NaN in pct\_pov\_below\_200 and pct\_singleparent\_hh  
4. fips 06095253000 (Highest) changed to NAs  
5. fips 06095980000 has NaN in pct\_pov\_below\_200 and pct\_singleparent\_hh  
6. fips 06095980000 (High) changed to NAs

##### The records below have poverty rate percentages, which is why they’re not changed to NAs to prevent them from not being counted even if they do not have pct\_single-parent\_ household\_hh. These records were categorized based on its index values.

1. fips 06001981900 has NA in pct\_singleparent\_hh
2. fips 06001981900 (High),
3. fips 06013351101 has NA in pct\_singleparent\_hh
4. fips 06013351101 (Mod),
5. fips 06013351102 has NA in pct\_singleparent\_hh
6. fips 06013351102 (Mod),
7. fips 06013351103 has NA in pct\_singleparent\_hh
8. fips 06013351103 (High),
9. fips 06075980300 has NA in pct\_singleparent\_hh
10. fips 06075980300 (High)

## Output: Index with Filters

table(data$category\_wo\_filters, data$category)

## < table of extent 0 x 0 >

merging <- full\_join(nan\_records, remaining)

## Joining, by = c("fips", "index", "DI\_Blk\_Lat\_POV30", "SPF30\_P30", "DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30")

#categories <- full\_join(remaining, categorize)  
#na\_records <- index %>% select("fips", "SPF\_GT\_30" , "pct\_singleparent\_hh\_2017.y") %>% filter(is.na(index$pct\_singleparent\_hh\_2017.y)) #8 NAs  
#index\_w\_filters <- full\_join(remaining, categorize)  
#write.csv(index\_w\_filters, "index\_w\_filters.csv")

### G. Overlays

#### Racial and Ethnic Composition Overlay

Data Source: ACS Census data 2010-2014  
Description: To analyze the distribution of racial and ethnic composition. I joined the shapefile using the ‘GEOID’ field to match it with the GEOID in the opportunity categories shapefile

race\_list <- c("fips", "CountyID.x.x", "TOTPOP.x.x", "total\_pop","index","total\_pop","white","black", "asian", "hispanic", "other", "county\_name","cbsa","divergence\_thresh", "DI\_Blk\_Lat\_POV30\_OR\_POV30\_SPF30")  
  
#filter\_race <- function(data, r\_list){  
# daf <- data %>% select(r\_list)  
# assign(daf, envir=.GlobalEnv)  
#}  
#source(here(filter\_race.R))  
#filter\_race(data=index\_filters, r\_list=race\_list)

#### Median Household Income

Data Source: American Community Survey (5-year-estimates)  
Table: B19013\_001 – MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2017 INFLATION-ADJUSTED DOLLARS)

#### Payday Lending Overlay

Data Source: ESRI Business Analyst  
Spreadsheet: OV\_YEAR\_Payday  
Description: 2017 Measure – Spatially join the payday lending in the bay area shape file to the 2014 census tract shape file with the opportunity categories to obtain the number of businesses per census tract. Then use the count of number of businesses per tract divided by the total count number of payday lending and credit businesses in the Bay Area to obtain the percentage.  
2018 Measure – Identify whether the column salevolume in the dataset has the volume of payday loan sales. Aggregate those sales and distribute them to tracts to identify the amount of sales in each neighborhood OR (if it’s possible to) identity where the highest percentage of interests (200-400%) that these payday loans are located and how many of them are in each census tracts.

#load(file="BA\_payday\_2018.RData")  
#proj4string(BA\_payday\_2018)

#### Subsidized Housing Overlay

Data Source: HUD subsidized housing projects  
Spreadsheet: OV\_Year\_SubHous  
Description:

• Data should be gathered through HUD instead of TCAC. Use the file obtained from HUD to create a point shapefile based on the lat and long for each (which is in the table).  
• This table has all subsidized housing projects in California; Use geoprocessing to clip the subsidized housing shapefile to Bay Area  
• Analysis of Projects and Units should be included in the map based on subsidized units available and the number of subsidized programs in the region.

#### Low population density Overlay

Data Source: Census Data  
Spreadsheet: OV\_Year\_LowDen  
Description: To analyze the density of the census tract and identify areas that are considered low density with 40 or more acres per person  
• Calculate the “area” of each tract in acres. Then I divided that by the number of people, and the results are in POP\_DEN field. All tracts which had a value of 40 or above were highlighted on the map with a specific symbology  
Example:  
Step 1: Create a new field, “Acres\_per” person for each tract > Calculate Geometry > selecting Area > Coordinate System: Use Coordinate System of the data frame: PCS: NAD 1983 StatePlane California III FIPS 0403 > Units: Acres [US] (ac) > OK  
Step 2: Then, create a new field titled, “POP\_DEN” in which the value would be “Acres\_per” person for each tract divided by the number of people in the tract > select the tracts that have the value of 40 or above