Calculating Gentrification Metrics Across UWIN sites

Mason Fidino

Questions we should answer

1. It seems like we need to subset this data a bit, as there are some sites with a very low density of people near the camera site in some cities (e.g., Salt Lake City). As our analysis is focused on calculating diversity metrics associated to people, how many people should be near a camera site for it to be included in the analysis? 100? 500? What is a cutoff that makes sense?

The data

The objective of this analysis is to determine how historical patterns of gentrification are associated to patterns of urban biodiversity. As such, we needed to compile census data from multiple years. To do so, I used the `tidycensus` package in `R` to query census data from the year 2000, 2010, 2015, and 2019. The 2000 data came from the 10-year decennial census whereas the remaining data come from the 5-year American Community Survey (ACS). The 10 year gap between 2000 and 2010 was because the 2005 5-year ACS data was not available. The 5-year ACS data was used because the 1-year ACS data did not contain estimates for smaller towns.

Across all of these years I compiled data on race, income, and educational attainment (Table 1) for all census tracts within 1 km of each camera trapping site. Following Freeman (2008), we reduced each of these variables down into relevant categories in order to calculate our gentrification metrics.

Table 1: Census variables gathered, and what the various categories were simplified to. See Table S2 for how income variables were reduced. * The ‘Other’ racial category was calculated by subtracting the other queried racial category from the Total population in each census tract.

data.type	variable	reduced.to
Education	No schooling completed	No high school diploma
	Nursery to 4th grade	No high school diploma
	5th and 6th grade	No high school diploma
	7th and 8th grade	No high school diploma
	9th grade	No high school diploma
	10th grade	No high school diploma
	11th grade	No high school diploma
	12th grade, no diploma	No high school diploma
	High school gradudate	High school diploma
	Some college, less than 1 year	Some college
	Some college, 1 or more years, no degree	Some college
	Associate degree	College graduate
	Bachelor’s degree	College graduate
	Master’s degree	Advanced college degree
	Professional school degree	Advanced college degree
	Doctorage degree	Advanced college degree
Income	Less than $10,000	See Table 2

data.type	variable	reduced.to
	$10,000 to $14,999	See Table 2
	$15,000 to $19,999	See Table 2
	$20,000 to $24,999	See Table 2
	$25,000 to $29,999	See Table 2
	$30,000 to $34,999	See Table 2
	$35,000 to $39,999	See Table 2
	$40,000 to $44,999	See Table 2
	$45,000 to $49,999	See Table 2
	$50,000 to $59,999	See Table 2
	$60,000 to $74,999	See Table 2
	$75,000 to $99,999	See Table 2
	$100,000 to $124,999	See Table 2
	$125,000 to $149,999	See Table 2
	$150,000 to $199,000	See Table 2
	$200,000 or more	See Table 2
Race	Total population	Other*
	White alone	White
	Black or African American alone	Black
	Asian alone	Asian
	Hispanic or Latino	Latino

Again, following Freeman (2008), the income categories were reduced to ‘poor’, ‘working class’, ‘middle class’ and ‘affluent’ based on the poverty line for a family of four for a given year (Table 2). Poor was defined as those making below the poverty line, working class was those who made above the poverty line, but less than twice the poverty line, middle class was those who made between two and three times the poverty line, and affluent were those who made more than four times above the poverty line. The poverty line was determined for each year from the poverty guidelines provided by the Office of the Assistant Secretary for Planning and Evaluation.

Table 2: The income classes used in this analysis. For a family of four, the poverty line was $17,050 in 2000, $22,050 in 2010, $25,100 in 2015, and $25,700 for 2019.

variable	2000	2010	2015	2020
Less than $10,000	poor	poor	poor	poor
$10,000 to $14,999	poor	poor	poor	poor
$15,000 to $19,999	poor	poor	poor	poor
$20,000 to $24,999	working class	poor	poor	poor
$25,000 to $29,999	working class	working class	working class	working class
$30,000 to $34,999	working class	working class	working class	working class
$35,000 to $39,999	middle class	working class	working class	working class
$40,000 to $44,999	middle class	working class	working class	working class
$45,000 to $49,999	middle class	middle class	working class	working class

variable	2000	2010	2015	2020
$50,000 to $59,999	middle class	middle class	middle class	middle class
$60,000 to $74,999	affluent	middle class	middle class	middle class
$75,000 to $99,999	affluent	affluent	affluent	affluent
$100,000 to $124,999	affluent	affluent	affluent	affluent
$125,000 to $149,999	affluent	affluent	affluent	affluent
$150,000 to $199,000	affluent	affluent	affluent	affluent
$200,000 or more	affluent	affluent	affluent	affluent

After querying all census tracts that were within 1 km of each camera trapping site we used areal interpolation to approximate the census metric within the 1 km buffers. Briefly, for s in $1, ..., S$ sites we calculated the total area of each census tract that intersected with site s ($\boldsymbol{a_s}$) as well as the area of the census tract that fell within the 1 km buffer ($\boldsymbol{b_s}$). We then created a weight by dividing $\boldsymbol{b_s}$ by $\boldsymbol{a_s}$, which represented the proportion of each census tract that fell within the buffer. Each census tracks weight was multiplied by their values associated to each census tract variable, and then we summed these weighted values across the census tracts that intersected with site s to approximate our metrics.

Calculating gentrification metrics: Education

After collecting and summarising the census data within 1 km of each camera, we converted all of the education categories to proportions. These data, like the income data, are ordinal cateories (there is an inherent order to the educational categories). As such, following Freeman (2008) we used applied the index of ordinal variation to these data (Kvalseth 1995) to these data, which is an ordinal measure of diversity. For k in $1, ..., K$ education categories this equation is

$$h_s = \frac{1}{k-1} \sum_{k=1}^{k-1} 4c_{s,k}(1 - c_{s,k})$$

where $c_{s,k}$ is the cumulative proportion of the total population at that level or lower at site s.

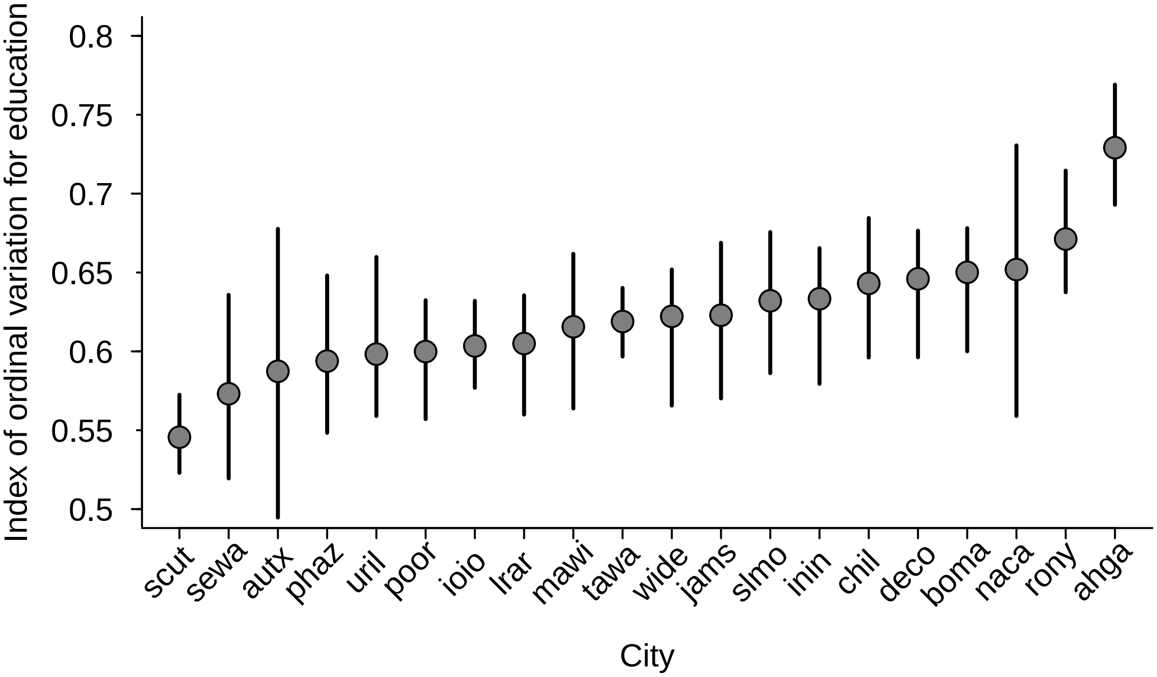
This index of ordinal variation, h_s, ranges between 0 and 1. For $K = 5$ categories $h_s = 1$ when the population is evenly split between the minimum and maximum ordinal categories, $h_s = 0.8$ when the proportion is evenly split among all categories, and $h_s = 0$ when the entire population is falls into a single category. Thus, a site would score higher if the people living near it were seperated between high school dropouts and people with advanced degrees, and would score low if all the people living near the site had advanced degrees. Essentially, this metric evaluates how diverse ordinal categories are from one another, with more positive values indicating more diversity.

Calculating the education metric in R

```r
# read in the data
edu <-read.csv("../data/cleaned_data/education.csv")

# change phaz2 to phaz
edu$city <- gsub("phaz2", "phaz", edu$city)

# make proportions
```

```r
edu$value <- edu$value / edu$total

# remove any sites that have no data
to_go <- unique(
  edu$site[is.na(edu$value)]
)

# just one site at the national capital
edu <- edu[-which(edu$site %in% to_go),]

# order categories
edu$variable <- factor(
  edu$variable,
  levels = c(
    "no hs diploma", "hs diploma", "some college",
    "college graduate", "advanced college degree"
  )
)

# reorder data
edu <- edu[order(edu$city, edu$year, edu$site, edu$variable),]

# provide probability for each site & year combo, get back the
#  ordinal score.
ordinal_diversity <- function(probs){
  k <- length(probs)
  ck <- cumsum(probs)
  tmp <- rep(NA, k-1)
  for(i in 1:length(tmp)){
    tmp[i] <- 4 * ck[i] * (1 - ck[i])
  }
  h_s <- (1 / (k-1)) * sum(tmp)
  return(h_s)
}

# calculate this metric across all cities, sites, and years
edu_d <- edu %>%
  dplyr::group_by(city, site, year) %>%
  dplyr::summarise(
    edu_div = ordinal_diversity(value),
    .groups = "drop_last"
)

# this ranges between 0.25 and 0.87 in this dataset. Let's
#  summarise a bit by city
edu_mean <- edu_d %>%
  dplyr::group_by(city) %>%
  dplyr::summarise(
    median = median(edu_div),
    lower = quantile(edu_div, probs = 0.25),
    upper = quantile(edu_div, probs = 0.75),
    .groups = "drop_last"
)
```

```
edu_mean <- edu_mean[order(edu_mean$median),]

# plot it out
par(xpd = NA)
bbplot::blank(xlim = c(1,20), ylim = c(0.5, 0.8), bty = 'l')
bbplot::axis_blank(1, 1:20, minor = FALSE)
bbplot::axis_blank(2,seq(0.5,0.8,0.1))
bbplot::axis_text(side = 2, las = 1, line = 0.75)
# x axis labels
text(
  x = c(1:20 - 0.25),
  y = 0.46,
  labels = edu_mean$city,
  srt = 45
)
for(i in 1:nrow(edu_mean)){
  lines(
    x = rep(i,2),
    y = c(
      edu_mean$lower[i],
      edu_mean$upper[i]
    ),
    lwd = 2
  )
}
# add the points
points(
  x = 1:20,
  y = edu_mean$median,
  pch = 21,
  bg = "gray50",
  cex = 1.5
)
bbplot::axis_text(
  "Index of ordinal variation for education",
  2, line = 2.8
)
bbplot::axis_text(
  "City",
  1,
  line = 2.8
)
```



Calculating gentrification metrics: Income

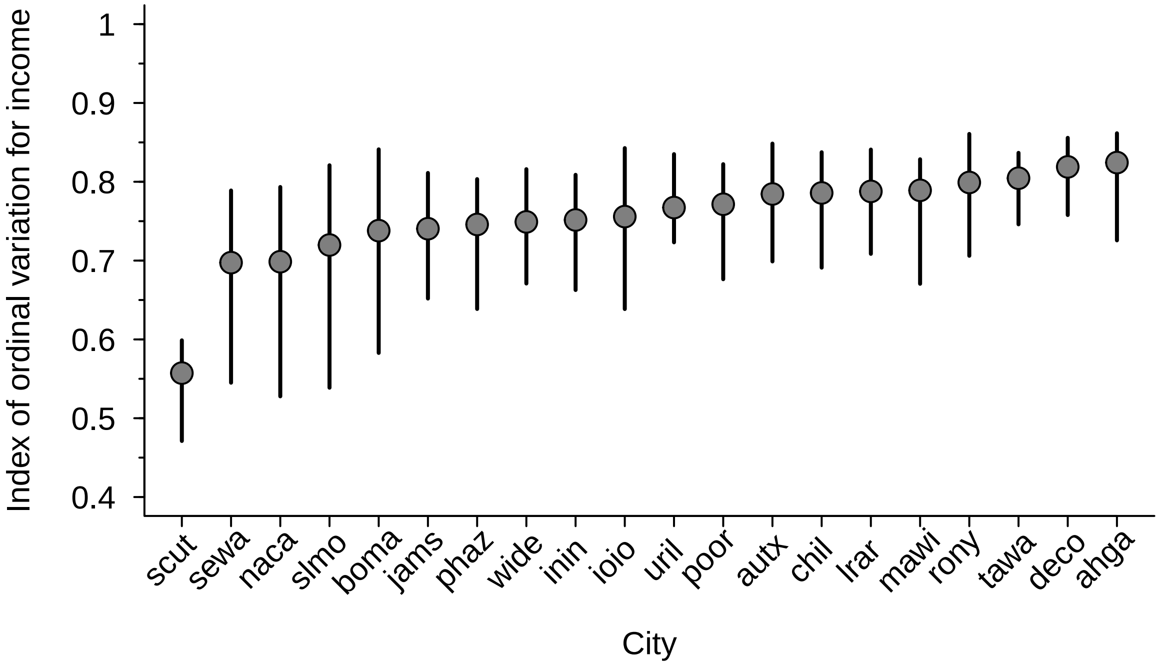
As income is also an ordinal variable, we can use the same exact ordinal diversity metric for these data as we did with education. For $K = 4$ categories, $h_s = 0.83$ when income groups are equally represented.

Calculating the income gentrification metric in R

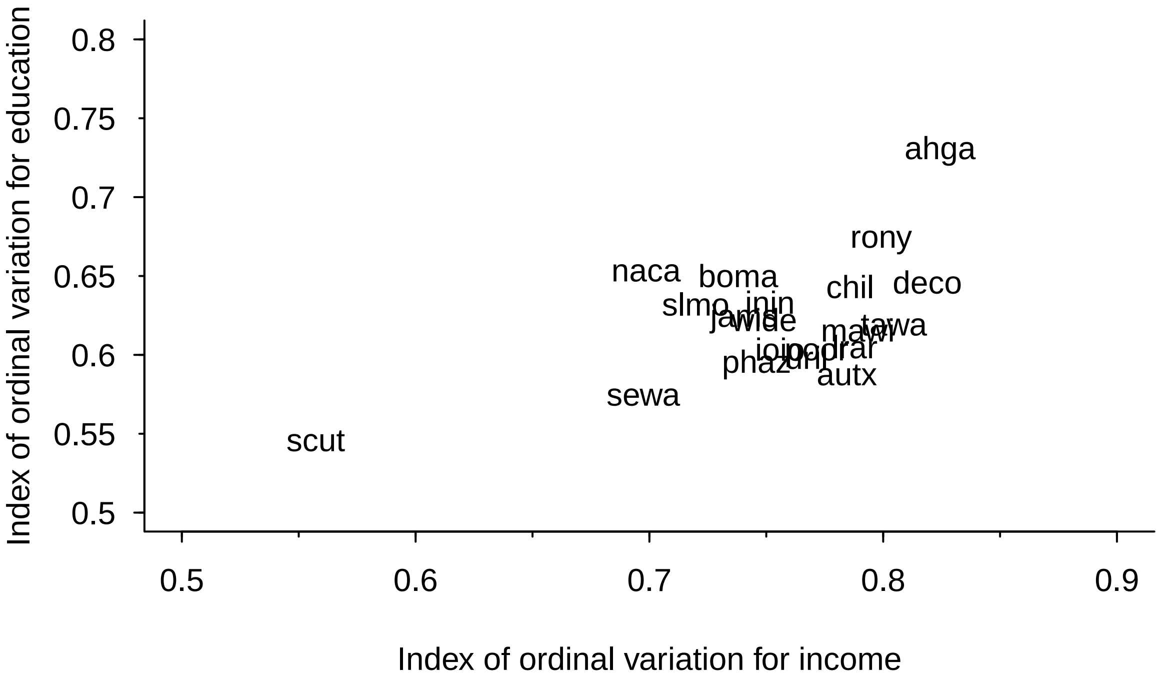
```
# read in the data
inc <-read.csv("../data/cleaned_data/income.csv")

# change phaz2 to phaz
inc$city <- gsub("phaz2", "phaz", inc$city)

# make proportions
inc$value <- inc$value / inc$total

# remove any sites that have no data
to_go <- unique(
  inc$site[is.na(inc$value)]
)

# just one site at the national capital
inc <- inc[-which(inc$site %in% to_go),]

# order categories
inc$variable <- factor(
```

```r
  inc$variable,
  levels = c(
   "poor", "working class", "middle class", "affluent"
  )
)

# reorder data
inc <- inc[order(inc$city, inc$year, inc$site, inc$variable),]


# calculate this metric across all cities, sites, and years
inc_d <- inc %>%
  dplyr::group_by(city, site, year) %>%
  dplyr::summarise(
    inc_div = ordinal_diversity(value),
    .groups = "drop_last"
)

# this ranges between 0.25 and 0.87 in this dataset. Let's
#  summarise a bit by city
inc_mean <- inc_d %>%
  dplyr::group_by(city) %>%
  dplyr::summarise(
    median = median(inc_div),
    lower = quantile(inc_div, probs = 0.25),
    upper = quantile(inc_div, probs = 0.75),
    .groups = "drop_last"
)

inc_mean <- inc_mean[order(inc_mean$median),]

# plot it out
par(xpd = NA)
bbplot::blank(xlim = c(1,20), ylim = c(0.4, 1), bty = 'l')
bbplot::axis_blank(1, 1:20, minor = FALSE)
bbplot::axis_blank(2,seq(0.4,1,0.1))
bbplot::axis_text(side = 2, las = 1, line = 0.75)
# x axis labels
text(
  x = c(1:20 - 0.25),
  y = 0.32,
  labels = inc_mean$city,
  srt = 45
)
for(i in 1:nrow(inc_mean)){
  lines(
    x = rep(i,2),
    y = c(
      inc_mean$lower[i],
      inc_mean$upper[i]
    ),
    lwd = 2
  )
```

```r
}
# add the points
points(
  x = 1:20,
  y = inc_mean$median,
  pch = 21,
  bg = "gray50",
  cex = 1.5
)
bbplot::axis_text(
  "Index of ordinal variation for income",
  2, line = 2.8
)
bbplot::axis_text(
  "City",
  1,
  line = 2.8
)
```



Now that we have two metrics, we can also plot them out against one another. From this plot, on average, Athens, Georgia (ahga) has the most diverse income and education metrics across their city whereas Salt Lake has the lowest. However, we may want to subset some of the city data as there are a decent number of sites with a very low density of people (21% of the dataset with less than 100 people within 1 square km of a camera site, and many of those camera sites are from Salt lake.)



Calculating gentrification metrics: Race/ethnic group

Unlike education or income, race/ethnic group is not an ordinal metric. So instead of using an ordinal measure of diversity, we used the entropy index to quantify census-tract level racial diversity (Freeman 2008). For r in $1, ..., R$ race/ethnic groups, let $p_{s,r}$ be the proportion of race/ethnic group r at site s and

$$q_s = \sum_{r=1}^{r=R} p_{s,r} \times \log(\frac{1}{p_{s,r}})$$

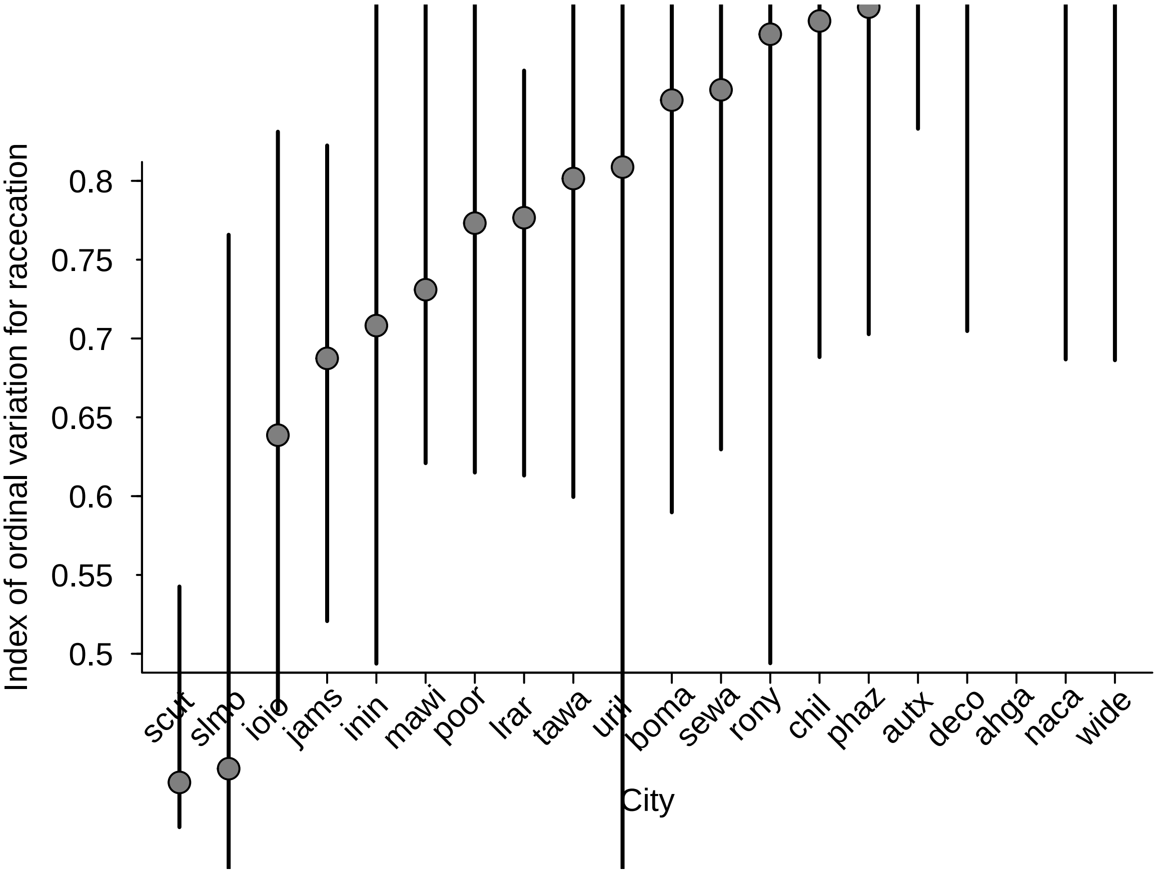
This metric gives higher values when race/ethnic groups are present in equal proportions and lower values when a single race/ethnic group is dominant. For $R = 5$ race/ethnic groups (white, Asian, Black, Latino, and other) the maximum value that q_s can take is 1.61, which happens when all groups are equally present. The minimum value for q_s is 0, which occurs when only one group is present.

Calculating the inc gentrification metric in R

```
# read in the data
race <- read.csv("../data/cleaned_data/race.csv")


# change phaz2 to phaz
race$city <- gsub("phaz2", "phaz", race$city)

# the 'other' category has not been made yet. calculate it for
#  each site and year
```

```r
race <- split(
  race,
  factor(
    paste0(race$site, "-", race$year)
  )
)
# Loop to calculate 'Other' category
for(i in 1:length(race)){
  race[[i]] <- rbind.data.frame(
    race[[i]],
    data.frame(
      city = race[[i]]$city[1],
      site = race[[i]]$site[1],
      variable = "Other",
      value = race[[i]]$total[1] - sum(race[[i]]$value),
      total = race[[i]]$total[1],
      year = race[[i]]$year[1]
    )
  )
}

# recombine back into one big data.frame
race <- do.call("rbind", race)



# make proportions
race$value <- race$value / race$total

# if any values are negative, we just need to set them to 0.
#  there are some that are REALLY close to zero but negative
#  due to rounding errors.
race$value[race$value<0] <- 0



# make categories a factor
race$variable <- factor(
  race$variable,
)

# reorder data
race <- race[order(race$city, race$year, race$site, race$variable),]

# provide probability for each site & year combo, get back the
#  ordinal score.

entropy <- function(probs){
  if(any(probs == 0)){
    probs <- probs[-which(probs == 0)]
  }
  q_s <- sum(probs * log((1/probs)))
  return(q_s)
```

```r
}


# calculate this metric across all cities, sites, and years
race_d <- race %>%
  dplyr::group_by(city, site, year) %>%
  dplyr::summarise(
    race_div = entropy(value),
    .groups = "drop_last"
)

# some sites are returning NA


# this ranges between 0.25 and 0.87 in this dataset. Let's
#  summarise a bit by city
race_mean <- race_d %>%
  dplyr::group_by(city) %>%
  dplyr::summarise(
    median = median(race_div),
    lower = quantile(race_div, probs = 0.25),
    upper = quantile(race_div, probs = 0.75),
    .groups = "drop_last"
)

race_mean <- race_mean[order(race_mean$median),]

# plot it out
par(xpd = NA)
bbplot::blank(xlim = c(1,20), ylim = c(0.5, 0.8), bty = 'l')
bbplot::axis_blank(1, 1:20, minor = FALSE)
bbplot::axis_blank(2,seq(0.5,0.8,0.1))
bbplot::axis_text(side = 2, las = 1, line = 0.75)
# x axis labels
text(
  x = c(1:20 - 0.25),
  y = 0.46,
  labels = race_mean$city,
  srt = 45
)
for(i in 1:nrow(race_mean)){
  lines(
    x = rep(i,2),
    y = c(
      race_mean$lower[i],
      race_mean$upper[i]
    ),
    lwd = 2
  )
}
# add the points
points(
  x = 1:20,
  y = race_mean$median,
  pch = 21,
  bg = "gray50",
  cex = 1.5
)
bbplot::axis_text(
  "Index of ordinal variation for racecation",
  2, line = 2.8
)
bbplot::axis_text(
  "City",
  1,
  line = 2.8
)
```


Index of ordinal variation for racecation
scut
slmo
ioio
jams
inin
mawi
poor
lrar
tawa
uril
boma
sewa
rony
chil
phaz
autx
deco
ahga
naca
wide
City