Final Project: Walkability and Public Health in the US

Lou Godmer, Kienan Battin, Divakar Mehta

April 17, 2023

Contents

Objective

The objective is to quantify the causal effect that the "walkability" of a region has on public health. The original data comes from two sources: 1. The U.S. Chronic Disease Indicators provides reported cases of a set of 124 indicators that are important to public health, and the geographic location of the case. 2. The Walkability Index quantifies every Census 2019 block group's relative "Walkability" as defined by the EPA based on characteristics such as easy walking access to public transit, jobs, stores and services. Quantifying the causal effect of walkability on public health can help policy makers understand how community planning measures that may improve or degrade the walkability of the region will impact public health.

The appendix of this document describes the pre-processing methodology that was used combine the two data sets to enable the quantitative analysis. Because the pre-processing methodology can take an hour or more to execute, the pre-processed data was exported. The beginning of this document imports the pre-processed data and the rest of the analysis is done based on the pre-processed data.

Load necessariy libararies

```r
rm(list=ls())

options(repos = list(CRAN="http://cran.rstudio.com/"))

if (!require('NHANES')) install.packages('NHANES')
library('openxlsx')

if (!require('ggplot2')) install.packages('ggplot2')
library('ggplot2')

if (!require('dplyr')) install.packages('dplyr')
library('dplyr')

if (!require('GGally')) install.packages('GGally')
library('GGally')

if (!require('tableone')) install.packages('tableone')
library(tableone)

if (!require('pROC')) install.packages('pROC')
library(pROC)

if (!require('tidycensus')) install.packages('tidycensus')
library(tidycensus)

if (!require('tigris')) install.packages('tigris')
library(tigris)

if (!require('sf')) install.packages('sf')
library(sf)

if (!require('stringr')) install.packages('stringr')
library(stringr)

if (!require('dplyr')) install.packages('dplyr')
library(dplyr)
```

Load the data

Download the data which has already undergone the pre-processing methodology described in the appendix. WARNING: this may take several minutes. To avoid unnecessary downloads, the commands are commented out. Un-comment and execute the commands to download the data.

```r
#download.file("https://walkabilityandhealth.blob.core.windows.net/walkabilityandhealth/disease_with_wa
#unzip("disease_with_walkability.zip", "disease_with_walkability.csv")
```

```r
disease_with_walkability <- read.csv("disease_with_walkability.csv")
```

Get familiar with the data

Descriptions of the fields in the dataset

The table below describes the fields that are used in this analysis

Data Set	Field Name	Field Description	Usage In This Analysis
Walkability	NatWalkInd	Walkability Index	Treatment Variable
Disease Indicators	LocationAbbr	US State or Territory Abbreviation	
Disease Indicators	LocationDesc	US State or Territory name	
Disease Indicators	DataSource	Origin of the disease indicator data	
Disease Indicators	Topic	Category of the disease information, i.e. "Asthma"	
Disease Indicators	Question	Brief description of the condition being measured	Dependent variable category
Disease Indicators	DataValueUnit	Unit of measurement for the response to "Question", i.e. "gallons"	
Disease Indicators	DataValueType	Type of measurement for the response to "Question", i.e. "mean"	
Disease Indicators	DataValueAlt	Numeric value of the response to "Question"	Dependent variable value
Disease Indicators	StratificationCatgory1	Category of demographic characteristic of the population being measured, i.e. Gender	Independent var (possible confounder)
Disease Indicators	Stratification1	Value of demographic characteristic of the population being measured, i.e. Female	Independent var (possible confounder)
Disease Indicators	GeoLocation	Longitude and latitude of the location where the data was collected	

Data Set	Field Name	Field Description	Usage In This Analysis
Disease Indicators	STATEFP	FIPS state code of the state of GeoLocation	
Disease Indicators	COUNTYFP	FIPS county code of the county of GeoLocation	
Disease Indicators	TRACTCE	FIPS tract code of the tract of GeoLocation	
Disease Indicators	BLKGRPCE	FIPS block code of the block group of GeoLocation	
Disease Indicators	GEOID	Full GEOID (state, county, tract, block group) of GeoLocation	
Walkability	CSA	"Combined Statistical Area" - grouping of adjacent metropolitan statistical areas that share social and economic ties	
Walkability	CSA_NAME	Friendly name of the CSA	
Walkability	CBSA	"Core Based Statistical Area" - functional region based around an urban center along with adjacent areas that are socioeconomically tied to the urban center by commuting	
Walkability	CBSA_NAME	Friendly name of the CBSA	
Walkability	CBSA_POP	Estimated population of the CBSA	
Walkability	CBSA_EMP	Total number of employees in the CBSA	
Walkability	CBSA_WRK	Total number of workers in the CBSA	
Walkability	AC_Total	Total area of land in square meters within the block group	
Walkability	AC_Water	Total area of land in square meters covered by water within the block group	
Walkability	AC_Land	Total area of land in square meters not covered by water within the block group	
Walkability	AC_Unpr	Total are of land in square meters classified as unproductive or unused within the block group	

Data Set	Field Name	Field Description	Usage In This Analysis
Walkability	TotPop	Total population within the block group	
Walkability	CountHU	Count of housing units in the block group	
Walkability	HH	Count of occupied housing units in the block group	
Walkability	P_WrkAge	Percentage of the population that is of working age (16 or older)	
Walkability	AutoOwn0	Households with zero automobiles	
Walkability	Pct_AO0	Percentage of households with zero automobiles	
Walkability	AutoOwn1	Households with one automobiles	
Walkability	Pct_AO1	Percentage of households with one automobiles	
Walkability	AutoOwn2p	Households with two or more automobiles	
Walkability	Pct_AO2p	Percentage of households with two or more automobiles	
Walkability	Workers	Population of workers (16 or older) in the block group	
Walkability	R_LowWageWk	Number of workers earning $1250/month or less (home location)	
Walkability	R_MedWageWk	Number of workers earning more than $1250/month and less than $3333/month (home location)	
Walkability	R_HiWageWk	Number of workers earning $3333/month or more (home location)	
Walkability	R_PCTLOWWAGE	Low wage workers as a percent of all workers in CBG (home location)	
Walkability	TotEmp	Total employment	
Walkability	E8_Ret	Retail jobs within a 8-tier employment classification scheme	
Walkability	E8_off	Office jobs within a 8-tier employment classification scheme	

Data Set	Field Name	Field Description	Usage In This Analysis
Walkability	E8_Ind	Industrial jobs within a 8-tier employment classification scheme	
Walkability	E8_Svc	Service jobs within a 8-tier employment classification scheme	
Walkability	E8_Ent	Entertainment jobs within a 8-tier employment classification scheme	
Walkability	E8_Ed	Education jobs within a 8-tier employment classification scheme	
Walkability	E8_Hlth Healthcare jobs within a 8-tier employment classification scheme		
Walkability	E8_Pub	Public administration jobs within a 8-tier employment classification scheme	
Walkability	E_LowWageWk	Number of workers earning $1250/month or less (work location)	
Walkability	E_MedWageWk	Number of workers earning more than $1250/month and less than $3333/month (work location)	
Walkability	E_HiWageWk	Number of workers earning $3333/month or more (work location)	
Walkability	E_PctLowWage	Low wage workers as a percent of all workers in CBG (work location)	
Walkability	D1A	Gross residential density (HU/acre) on unprotected land	
Walkability	D1B	Gross population density (people/acre) on unprocted land	
Walkability	D1C	Gross employment density (jobs/acre) on unprotected land	
Walkability	D1C8_RET	Gross retail (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1C8_OFF	Gross office (8-tier) employment density (jobs/acre) on unprotected land	

Data Set	Field Name	Field Description	Usage In This Analysis
Walkability	D1C8_IND	Gross industrial (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1C8_SVC	Gross service (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1C8_ENT	Gross entertainment (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1C8_ED	Gross education (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1C8_HLTH	Gross healthcare (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1C8_PUB	Gross public administration (8-tier) employment density (jobs/acre) on unprotected land	
Walkability	D1D	Gross activity density (HU + employment / acre) on unprotected land	
Walkability	D2A_JPHH	Jobs per housing unit	
Walkability	D2B_E8MIX	8-tier employment entropy	
Walkability	D2B_E8MIXA	8-tier employment entropy, denominator set to the static 8 employment types in the CBG	
Walkability	D2C_TRPMX2	Employment and household entropy (excluding industrial jobs), based on trip production and attraction	
Walkability	D2C_TRIPEQ	Trip production and trip attractions equilibrium index (closer to 1 = more balance)	
Walkability	D2R_JOBPOP	Deviation of CBG jobs/population ratio from regional average jobs/pop ratio	

Data Set	Field Name	Field Description	Usage In This Analysis
Walkability	D2R_WRKEMP	Household workers per job	
Walkability	D2A_WORKEMP	Deviation of CBG ratio of household workers/job from regional average ratio of household workers/ob	
Walkability	D2C_WREMLX	Household worker per job equilibrium index (closer to one = more balanced)	
Walkability	D4A	Distance from population weighted centroid to nearest transit stop, meters	
Walkability	D4B025	Proportion of CBG employment within 1/4 mile of fixed guideway transit stop	
Walkability	D4B050	Proportion of CBG employment within 1/2 mile of fixed guideway transit stop	
Walkability	D4C	Transit service frequency. (Afternoon peak period transit departure within 0.25 miles)	
Walkability	D4D	Peak pm transit departure within 0.25 miles of CBG, per square mile	
Walkability	D5AR	Jobs within a 45 minute drive (weighted)	
Walkability	D5AE	Working-age population within 45 min. drive (weighted)	
Walkability	D5BR	Jobs within 45 min. transit commute (weighted)	
Walkability	D5BE	Working-age population within 45 min. transit commute (weighted)	
Walkability	D5CR	Job accessibility (D5ar) as proportion of total regional job accessibility	
Walkability	D5CRI	Regional centrality index (auto) - D5cr divided by max D5cr in metro region (CBSA)	

Data Set	Field Name	Field Description	Usage In This Analysis
Walkability	D5CE	Accessibility to working-age populatin (D5ae) as proportion of total regional accessibility	
Walkability	D5CEI	Regional centrality index (auto) - D5ce divided by max D5ce in metro region (CBSA)	
Walkability	D5DR	Job accessibility by transit (D5br) as proportion of total regional job accessibility by transit	
Walkability	D5DRI	Regional centrality index (transit) - D5dr divided by max D5dr in metro region (CBSA)	
Walkability	D5DE	Accessibility to working-age populatin by transit (D5be) as proportion of total regional accessibility	
Walkability	D5DEI	Regional centrality index (transit) - D5de divided by max D5de in metro region (CBSA)	

```
str(disease_with_walkability)
```

```
## ’data.frame':    888329 obs. of  155 variables:
##  $ X                         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ YearStart                 : int  2014 2018 2018 2017 2010 2015 2013 2013 2017 2010 ...
##  $ YearEnd                   : int  2014 2018 2018 2017 2010 2015 2013 2013 2017 2010 ...
##  $ LocationAbbr              : chr  "AR" "CO" "DC" "GA" ...
##  $ LocationDesc              : chr  "Arkansas" "Colorado" "District of Columbia" "Georgia" ...
##  $ DataSource                : chr  "SEDD; SID" "SEDD; SID" "SEDD; SID" "SEDD; SID" ...
##  $ Topic                     : chr  "Asthma" "Asthma" "Asthma" "Asthma" ...
##  $ Question                  : chr  "Hospitalizations for asthma" "Hospitalizations for asthma" "Hospi
##  $ Response                  : logi  NA NA NA NA NA NA ...
##  $ DataValueUnit             : chr  "" "" "" "" ...
##  $ DataValueType             : chr  "Number" "Number" "Number" "Number" ...
##  $ DataValue                 : chr  "916" "2227" "708" "3520" ...
##  $ DataValueAlt              : num  916 2227 708 3520 123 ...
##  $ DataValueFootnoteSymbol   : chr  "" "" "" "" ...
##  $ DatavalueFootnote         : chr  "" "" "" "" ...
##  $ LowConfidenceLimit        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ HighConfidenceLimit       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ StratificationCategory1   : chr  "Gender" "Overall" "Overall" "Gender" ...
##  $ Stratification1           : chr  "Male" "Overall" "Overall" "Female" ...
##  $ StratificationCategory2   : logi  NA NA NA NA NA NA ...
##  $ Stratification2          : logi  NA NA NA NA NA NA ...
##  $ StratificationCategory3  : logi  NA NA NA NA NA NA ...
##  $ Stratification3          : logi  NA NA NA NA NA NA ...
##  $ GeoLocation              : chr  "POINT (-92.27449074299966 34.74865012400045)" "POINT (-106.13361
##  $ ResponseID               : logi  NA NA NA NA NA NA ...
##  $ LocationID               : int  5 8 11 13 26 30 41 72 72 55 ...
##  $ TopicID                  : chr  "AST" "AST" "AST" "AST" ...
##  $ QuestionID               : chr  "AST3_1" "AST3_1" "AST3_1" "AST3_1" ...
##  $ DataValueTypeID          : chr  "NMBR" "NMBR" "NMBR" "NMBR" ...
##  $ StratificationCategoryID1: chr  "GENDER" "OVERALL" "OVERALL" "GENDER" ...
##  $ StratificationID1        : chr  "GENM" "OVR" "OVR" "GENF" ...
##  $ StratificationCategoryID2: logi  NA NA NA NA NA NA ...
##  $ StratificationID2        : logi  NA NA NA NA NA NA ...
##  $ StratificationCategoryID3: logi  NA NA NA NA NA NA ...
##  $ StratificationID3        : logi  NA NA NA NA NA NA ...
##  $ lat                      : num  34.7 38.8 38.9 32.8 44.7 ...
##  $ long                     : num  -92.3 -106.1 -77 -83.6 -84.7 ...
##  $ STATEFP                  : int  5 8 11 13 26 30 41 72 72 55 ...
##  $ COUNTYFP                 : int  119 15 1 21 39 27 69 73 73 141 ...
##  $ TRACTCE                  : int  4400 404 5303 13701 960200 30201 960100 956305 956305 11000 ...
##  $ BLKGRPCE                 : int  1 1 1 1 2 1 2 1 1 5 ...
##  $ GEOID                    : num  5.12e+10 8.02e+10 1.10e+11 1.30e+11 2.60e+11 ...
##  $ OBJECTID                 : int  30558 NA NA NA 122446 NA 183638 NA NA 216412 ...
##  $ GEOID10                  : num  5.12e+10 NA NA NA 2.60e+11 ...
##  $ GEOID20                  : num  5.12e+10 NA NA NA 2.60e+11 ...
##  $ CSA                      : int  340 NA NA NA NA NA NA NA NA 554 ...
##  $ CSA_Name                 : chr  "Little Rock-North Little Rock, AR" NA NA NA ...
##  $ CBSA                     : int  30780 NA NA NA NA NA NA NA NA 49220 ...
##  $ CBSA_Name                : chr  "Little Rock-North Little Rock-Conway, AR" NA NA NA ...
##  $ CBSA_POP                 : int  734502 NA NA NA 0 NA 0 NA NA 73274 ...
##  $ CBSA_EMP                 : int  346204 NA NA NA 0 NA 0 NA NA 39593 ...
##  $ CBSA_WRK                 : int  315683 NA NA NA 0 NA 0 NA NA 38537 ...
##  $ Ac_Total                 : num  427 NA NA NA 20496 ...
##  $ Ac_Water                 : num  28.8 NA NA NA 230.6 ...
##  $ Ac_Land                  : num  398 NA NA NA 20266 ...
##  $ Ac_Unpr                  : num  393 NA NA NA 6395 ...
##  $ TotPop                   : int  1228 NA NA NA 1879 NA 756 NA NA 648 ...
##  $ CountHU                  : int  1260 NA NA NA 857 NA 596 NA NA 237 ...
##  $ HH                       : int  948 NA NA NA 672 NA 355 NA NA 237 ...
##  $ P_WrkAge                 : num  0.816 NA NA NA 0.591 NA 0.526 NA NA 0.727 ...
##  $ AutoOwn0                 : int  226 NA NA NA 26 NA 0 NA NA 33 ...
##  $ Pct_AO0                  : num  0.2384 NA NA NA 0.0387 ...
##  $ AutoOwn1                 : int  527 NA NA NA 146 NA 88 NA NA 107 ...
##  $ Pct_AO1                  : num  0.556 NA NA NA 0.217 ...
##  $ AutoOwn2p                : int  195 NA NA NA 500 NA 267 NA NA 97 ...
##  $ Pct_AO2p                 : num  0.206 NA NA NA 0.744 ...
##  $ Workers                  : int  719 NA NA NA 555 NA 279 NA NA 431 ...
##  $ R_LowWageWk              : int  154 NA NA NA 143 NA 107 NA NA 109 ...
##  $ R_MedWageWk              : int  223 NA NA NA 231 NA 106 NA NA 171 ...
##  $ R_HiWageWk               : int  342 NA NA NA 181 NA 66 NA NA 151 ...
##  $ R_PCTLOWWAGE             : num  0.214 NA NA NA 0.258 ...
##  $ TotEmp                   : int  21225 NA NA NA 677 NA 155 NA NA 947 ...
##  $ E5_Ret                   : int  251 NA NA NA 77 NA 18 NA NA 15 ...
##  $ E5_Off                   : int  11152 NA NA NA 197 NA 16 NA NA 694 ...
##  $ E5_Ind                     : int  1966 NA NA NA 94 NA 58 NA NA 3 ...
##  $ E5_Svc                     : int  5237 NA NA NA 230 NA 59 NA NA 105 ...
##  $ E5_Ent                     : int  2619 NA NA NA 79 NA 4 NA NA 130 ...
##  $ E8_Ret                     : int  251 NA NA NA 77 NA 18 NA NA 15 ...
##  $ E8_off                     : int  5546 NA NA NA 18 NA 1 NA NA 96 ...
##  $ E8_Ind                     : int  1966 NA NA NA 94 NA 58 NA NA 3 ...
##  $ E8_Svc                     : int  4324 NA NA NA 101 NA 7 NA NA 51 ...
##  $ E8_Ent                     : int  2619 NA NA NA 79 NA 4 NA NA 130 ...
##  $ E8_Ed                      : int  186 NA NA NA 71 NA 48 NA NA 52 ...
##  $ E8_Hlth                    : int  727 NA NA NA 58 NA 4 NA NA 2 ...
##  $ E8_Pub                     : int  5606 NA NA NA 179 NA 15 NA NA 598 ...
##  $ E_LowWageWk                : int  3162 NA NA NA 222 NA 57 NA NA 204 ...
##  $ E_MedWageWk                : int  6910 NA NA NA 226 NA 61 NA NA 398 ...
##  $ E_HiWageWk                 : int  11153 NA NA NA 229 NA 37 NA NA 345 ...
##  $ E_PctLowWage               : num  0.149 NA NA NA 0.328 ...
##  $ D1A                        : num  3.21 NA NA NA 0.134 ...
##  $ D1B                        : num  3.129 NA NA NA 0.294 ...
##  $ D1C                        : num  54.076 NA NA NA 0.106 ...
##  $ D1C5_RET                   : num  0.639 NA NA NA 0.012 ...
##  $ D1C5_OFF                   : num  28.4123 NA NA NA 0.0308 ...
##  $ D1C5_IND                   : num  5.0088 NA NA NA 0.0147 ...
##  $ D1C5_SVC                   : num  13.342 NA NA NA 0.036 ...
##  $ D1C5_ENT                   : num  6.6725 NA NA NA 0.0124 ...
##  $ D1C8_RET                   : num  0.639 NA NA NA 0.012 ...
##  $ D1C8_OFF                   : num  14.12973 NA NA NA 0.00281 ...
##   [list output truncated]
```

The “Question” and “Response” fields contain data about an individuals response to various questions about disease indicators such as whether they have been hospitalized for asthma.

TODO: Insert rest of paper here

Appendix

Original data pre-processing methodology

As described in the objective section, the original data came from two sources. The disease indicators data contains location information in the form of latitude and longitude. The walkability data contains location information in the form of Federal census location codes (FIPS codes). The pre-processing technique below was used to convert the latitude and longitude to FIPS codes, and then perform a join operation utilizing the FIPS codes. The resulting data is the original disease indicators data, augmented with the walkability information for the location corresponding to the original latitude and longitude.

In other words, for every row in the disease indicators data set, the corresponding walkability information for the region was added to that row. All of the commands are commented out to prevent them from being executed on knit since they take a long time to run.

```
#download.file("https://edg.epa.gov/EPADataCommons/public/OA/EPA_SmartLocationDatabase_V3_Jan_2021_Fina
#download.file("https://data.cdc.gov/api/views/g4ie-h725/rows.csv?accessType=DOWNLOAD", destfile="disea
```

Download the raw data

```
#walkability <- read.csv("walkability.csv")
## some of the disease data has no GeoLocation, which we cannot use for our analysis, so filter those o
#disease <- filter(read.csv("diseaseindicators.csv"), GeoLocation != "")
```

Load the data into R

```
## Extract the latitude and longitude values from the GeoLocation column using str_extract_all()
#geo_df <- str_extract_all(disease$GeoLocation, "-?[0-9]+\\.[0-9]+")

## Convert the extracted values to numeric and assign them to the corresponding latitude and longitude
#disease$lat <- as.numeric(sapply(geo_df, function(x) x[2]))
#disease$long <- as.numeric(sapply(geo_df, function(x) x[1]))
```

Extract the latitude and longitude into separate columns

Fetch the geographic information required to map latitude and logitude to FIPS blocks The tigris library provides a function "block_groups" which returns geographic information about every FIPS block. This geographic information can be used to convert latitude and longitude to FIPS block. The following code downloads all of the block_groups for every block in the walkability data set.

```
## create data frame for block_groups data
#allblockgroups <- data.frame(matrix(ncol=6, nrow=0))
#colnames(allblockgroups) <- c('STATEFP', 'COUNTYFP', 'TRACTCE', 'BLKGRPCE', 'GEOID', 'geometry')

## get block geography data for each state in the walkability dataset
#stateCodes <- data.frame(unique(walkability$STATEFP))
#for (i in 1:nrow(stateCodes)) {
#  stateCode=stateCodes[[1]][i]
#  counties = distinct(filter(walkability, STATEFP == stateCode), COUNTYFP)$COUNTYFP
#  new_blocks <- block_groups(state=stateCodes[[1]][i], counties) %>%
#    select(STATEFP, COUNTYFP, TRACTCE, BLKGRPCE, GEOID, geometry)
#  allblockgroups <- rbind(allblockgroups, new_blocks)
#}
```

```
#my_points <- data.frame(
#  x = disease$lat,
#  y = disease$long
#) %>%
#  st_as_sf(coords = c("y", "x"),
#     crs = st_crs(allblockgroups))

#my_points_blocks <- st_join(my_points, allblockgroups)
#disease$STATEFP = as.integer(my_points_blocks$STATEFP)
#disease$COUNTYFP = as.integer(my_points_blocks$COUNTYFP)
#disease$TRACTCE = as.integer(my_points_blocks$TRACTCE)
```

```
#disease$BLKGRPCE = as.integer(my_points_blocks$BLKGRPCE)
#disease$GEOID = as.numeric(my_points_blocks$GEOID)
```

Use block geographies to convert longitude and latitude to FIPS blocks

```


# Join the disease data with the walkability data
#disease_with_walkability <- left_join(disease, walkability,
#                                  by = c("STATEFP", "COUNTYFP", "TRACTCE", "BLKGRPCE"))
```

Join the disease indicators and walkability data sets based on FIPS blocks

```


#write.csv(disease_with_walkability, file = "disease_with_walkability.csv")
```

Export the joined data to be used for further processing later.