CrossMark

A Reproducible Framework for Visualizing Demographic Distance Profiles in US Metropolitan Areas

Kyle E. Walker[1] (ORCID)

Abstract Distance profiles have long been used in urban demography to explore how demographic characteristics of metropolitan areas vary by distance from their urban cores. Distance profile visualizations graphically illustrate these relationships and are useful in exploratory demographic data analysis of urban areas. The purpose of this article is to demonstrate how to build distance profile visualizations reproducibly within R, a free and open-source programming language and data analysis environment. The approach to distance profile visualization in this article involves the graphical display of a smoothed relationship between the location quotient of a demographic group for a metropolitan Census tract and the distance between the tract centroid and its respective urban core. Data acquisition, analysis, and visualization are all handled in R. The tidycensus, sf, and ggplot2 R packages are featured in this framework. Distance profile visualizations for educational attainment are used as illustrative examples, and reveal how the geography of metropolitan educational attainment varies both over time and across different types of metropolitan areas.

Keywords Distance profiles · Data visualization · R · Census · American Community Survey

---

✉ Kyle E. Walker
  kyle.walker@tcu.edu

[1] Department of Geography, Texas Christian University, TCU Box 297280, Fort Worth, TX 76129, USA

⌀ Springer

1 Introduction

Ecological models of cities, pioneered by the Chicago School of Sociology in the 1920s, have figured prominently in scholarship and popular understandings of the demographic structure of American metropolitan areas over the past century. In the Chicago School model, cities are represented as concentric zones in which geographical location within the metropolis is linked with social position. For example, the Chicago scholars posited that low-income areas tended to be located near to urban cores, whereas affluence increased with distance from the central business district. While the Chicago model of urban form has been re-formulated and critiqued extensively over the past 100 years, the relationship between social class and metropolitan location still resonates in stereotypical depictions of "the inner city" and "suburbia".

Given this theoretical framework, scholars have long been interested in studying where different demographic groups reside in metropolitan areas by distance from the city center. A common tool for exploring this is the *distance profile visualization* (Wilson et al. 2012), a technique that illustrates a modeled or measured relationship between group concentration and relative distance from the urban core. While this technique itself is not new, recent advances in scientific computing software such as the R programming language have made the development of such exploratory visualizations much more accessible.

The purpose of this article is to introduce a reproducible framework for computing and visualizing demographic distance profiles for US metropolitan areas, using data from the US Census Bureau. In this framework, all data acquisition, modeling, and visualization is handled within the R environment, meaning that no external GIS software or data preparation is necessary. The framework is illustrated with the example of educational attainment for large metropolitan areas in the United States.

2 Distance Profiles of American Cities

How the morphology and demography of neighborhoods varies by distance from the city center has long been of interest to urban scholars across the social sciences. While these studies vary widely in topics and in mathematical approaches, they are united by a shared belief that these within-metropolitan variations reveal something salient about the structure and characteristics of cities. In turn, I characterize this literature as concerned with the *distance profiles* of neighborhood characteristics, which can be represented graphically through *distance profile visualization*.

A prominent example is the *density gradient*, which has been used to explain how metropolitan areas transition from high-density neighborhoods near to the urban core to low-density neighborhoods on the fringes. Clark (1951) argues that urban population densities are characterized by *exponential decline*, represented by the equation $y = Ae^{-bx}$, where x is the distance from the city center and y is the neighborhood population density. A, then, represents the density at the city center, and b governs the degree to which density drops off from the urban core to the

fringe. To illustrate the application of this formula, Clark uses distance profile visualization to show how density falls with distance from the city center in cities around the world. Follow-up research (e.g. Berry et al. 1963; Lahiri et al. 1989; Alperovich and Deutsch 1992) has proposed alternative mathematical formulas for modeling the density gradient, and conducted additional comparative work across different types of cities.

The concept of the distance profile is also represented in the economics and economic geography literatures. Eberts (1981) and McMillen and Singell (1992), for example, write of the intraurban *wage gradient*, which suggests that wages decline with distance from the city center. This literature is accompanied by research on *rent*, *home value*, and *income gradients*, which assess and model the relationships between distance from the urban core and these economic variables. In the 1960s, scholars such as Alonso (1964) and Muth (1969) propose that prices fall with distance from the central business district. More recent research has sought to account for how the polycentric nature of modern metropolitan areas influences such gradients. Hackworth (2005) compares these gradients across American metropolitan areas to investigate the applicability of mono- and polycentric models of urban form in different cities. Other representative scholarship in this area includes Gong et al. (2016) who model such gradients for Chinese cities, and Albouy and Lue (2015) who consider relationships between wage and rent gradients when accounting for factors such as household characteristics and commuting.

Hackworth's approach is notable here as it uses visualization to graphically represent these gradients across large American metropolitan areas. The visuals are bar charts with bars for 10 km distance bands from the urban core, and where the height of the bars represents the percentage of the metropolitan average. In turn, these charts can be characterized as *distance profile visualizations* as they allow for visual comparison of such trends across the cities in the study.

Other recent scholarship has used visualization to represent the dynamics of distance profiles. Notable examples include the work of Estiri et al. (2015) and Estiri and Krause (2016), whose papers examine the relationships between residential location within metropolitan areas and the life course. In both papers, they find that younger households are more likely to reside near to the urban core, whereas older households are more prevalent on the fringes. Their approach involves the calculation of a location quotient for different age cohorts to measure relative cohort concentration at the Census block level, and the fitting and graphing of LOESS models to visualize the local relationships between cohort concentration and distance from the urban core. Walker (2018a, b) adopts a similar approach to study the geography of racial and ethnic diversity across American cities. In this paper, Walker visualizes *diversity gradients* to explore how Census tract-level racial diversity—as measured by the entropy index—varies by distance from their respective metropolitan urban cores, also with LOESS smoothing.

Distance profile visualization is also used in some prominent public reports. The US Census Bureau (Wilson et al. 2012) uses distance profile visualization to compare population densities and population distributions by distance from the city hall of the major metropolitan city for the New York, Los Angeles, Dallas-Fort Worth, and Miami areas. Juday (2015a) extends this by producing distance profile

visualization for topics such as age, race, and education across several American metropolitan areas, and compares the distance profiles of these metros between 1990 and 2012. Accompanying Juday's report is an interactive website that allows visitors to explore these graphs for many large American cities (Juday 2015b), which were reprinted in *The New York Times* and *The Washington Post* (Brown and Shapiro 2015; Edsall 2015).

The academic and public-facing usage of distance profile visualization illustrates the utility of this trend for exploring population distributions in metropolitan areas, and for stimulating hypothesis development to inform urban modeling. This article presents a reproducible framework for producing distance profile visualizations for any variable from the decennial US Census or American Community Survey. To do so, the article will illustrate the utility of two new R packages for spatial data acquisition and analysis: **tidycensus**, which enables R users to obtain decennial Census and/or American Community Survey data linked with feature geometry in a single function call (Walker 2018a, b), and **sf**, a package that represents spatial data in R like data frames, with feature geometry in a list-column (Pebesma 2017). The R language is particularly well-suited for demographic data analysis and visualization. Examples include Sparks (2014), who illustrates how to compute measures of residential segregation using Census data obtained within the R environment, and Walker (2016a, b) who demonstrates the use of R to obtain international demographic data and create interactive exploratory visualizations such as population pyramids. Like the workflow used in these aforementioned articles, this article shows how an *entire* process of Census data acquisition, spatial analysis, and visualization to take place within an R script. To illustrate this framework, this article examines the distribution of population subgroups by educational attainment within metropolitan areas, with attention to how their residential locations vary across metros. However, the framework can be extended to any other topic of interest.

3 Data and Methods

The analysis in this paper follows the methodology employed by Estiri and Krause (2016). Their study measures group concentration as the group *location quotient* (*LQ*) for a tract's corresponding metropolitan area. *LQ* for a Census tract *i* in a given metropolitan area *m* is computed as follows:

$$LQ = \frac{G_i/T_i}{G_m/T_m}$$

where *G* is the group population and *T* is the total population.

"Metropolitan areas" in this approach are represented by Census Bureau core-based statistical area boundaries. In this definition, metropolitan areas are specified as collections of counties surrounding a core urban area of at least 50,000 people. Counties beyond the core urban area are included within the metropolitan area if

they a have "a high degree of social and economic integration" with the core area, measured by commuting patterns (Office of Management and Budget 2017).

Tract location quotients are then used as inputs in a locally weighted regression model (LOESS) which represents the smoothed relationship between group concentration and either distance from the urban core in a given metropolitan area. By convention, LOESS models are estimated using weighted least squares with a tricubic weighting scheme computed as $1 - ((\frac{d}{D})^3)^3$, where d represents the distance between points and D is the neighborhood maximum. The size of the neighborhood is governed by a span parameter α, which can be modified by the user to control the degree of smoothing on the plot. The examples in this article use a span parameter of 0.3, meaning that LOESS estimates account for the nearest 30 percent of observations in variable space; however, users can modify this as needed when reproducing the results if a smoother or more granular fit is desired.

Data for this example come from the 2012–2016 American Community Survey's *Data Profiles*, which include commonly requested socio-demographic information from the ACS (US Census Bureau 2017). Data acquisition is handled using the R package **tidycensus** (Walker 2018a, b). tidycensus is a joint interface to the US Census Bureau's decennial Census and five-year American Community Survey application programming interfaces (APIs), and its repository of TIGER/Line and cartographic boundary files. Using functions in tidycensus, R programmers can request data from the Census or ACS using their corresponding APIs, returning **tidy** data frames ready for use with the **tidyverse** suite of packages in R (Wickham 2014; Wickham and Grolemund 2017). Following Wickham (2014, 4), in a tidy dataset "1. Each variable forms a column. 2. Each observation forms a row. 3. Each type of observational unit forms a table." tidycensus aims to align with this tidy data model as much as possible, with some small exceptions given the unique characteristics of US Census Bureau data. By default, tidycensus returns data as **tibbles**, which are modified versions of standard R data frames optimized for better interactive display.

The returned tibble includes a *GEOID* column, representing the unique Census ID code of each observational unit; *NAME*, which is a descriptive name of the unit; and *variable*, which is the Census ID code for the requested variable or variables. For decennial Census data, a *value* column is returned, representing the value of the requested variable for a given enumeration unit; for ACS data, *estimate* and *moe* columns are returned representing the estimate for the requested variable and the margin of error around that estimate. In turn, each row in a tidycensus data frame, by default, represents a unique enumeration unit-Census/ACS variable combination.

Additionally, if the tidycensus user specifies *geometry = TRUE* in a tidycensus function call for core enumeration units (states, counties, tracts, block groups, blocks, and ZCTAs for the ACS), the function will return a tibble of Census or ACS data along with a list-column of simple feature geometry named *geometry*. As the returned object is also an object of class *sf*, R users can map the returned data or perform spatial analysis. Feature geometry from the US Census Bureau is retrieved using the R **tigris** package (Walker 2016a, b), by default using the Census Bureau's cartographic boundary shapefiles.

The illustrative example used in this article for building distance profile visualizations is *educational attainment*. Historically, an application of the Chicago model to educational attainment within metropolitan areas would suggest that educational attainment should rise with distance from the urban core (Schnore 1963; Taggart 1971). However, there is a wealth of evidence that suggests a more complex relationship. Using 1960 data, Schnore (1963) finds that in smaller and newly urbanized regions, city socioeconomic status may outstrip that of the surrounding suburbs. Such nuanced relationships between residential location and educational attainment are also found in studies of spatial assimilation; Alba and Logan (1991), for example, find that while an increase in educational attainment is associated with suburban residential location for most groups, the inverse is true for non-Hispanic whites and for Japanese. More recent scholarship suggests a growing relationship between central city location and high educational attainment; Sander (2006) finds that central cities of large metropolitan areas in the US have high levels of educational attainment, but also high school dropout rates; Walker (2017) also finds that migrants to large metropolitan areas with bachelor's or graduate degrees are far more likely to move near to the urban core than migrants with high school diplomas as their highest level of educational achievement.

This literature suggests that educational attainment is an appropriate variable for demonstrating the utility of distance profile visualization as an exploratory method. As the relationship suggested in the literature between educational attainment and urban/suburban residential location is not a clear linear gradient, it is worthwhile to represent it using a smoother like LOESS. Further, the framework developed in this article will demonstrate how to build such visualizations to compare distance profiles across metropolitan areas, capturing metropolitan-specific variations; make comparisons of distance profile visualizations over time; and explore model outliers using geographic visualization and tools for linked brushing of charts and maps.

4 Building Distance Profile Visualizations in R

To get started, the user should have installed R (version 3.3 or higher) and should install the required packages from CRAN with the *install.packages()* command. tidycensus, tigris, and sf are required, along with the **tidyverse** package which loads a suite of packages for data wrangling and visualization in R that will be used in this workflow. Once the required packages are installed, the user should load the required packages and set some environment options. To get data from the US Census Bureau API, an API key is required; this can be obtained from http://api.census.gov/data/key_signup.html and set in a user's R session with *census_api_key()*. The command *options(tigris_class = "sf")* will instruct tigris to load Census geometry as simple features; *options(tigris_use_cache = TRUE)* is not strictly necessary, but will cache Census shapefiles on a user's machine for faster future access.

```
library(tidycensus)
library(tidyverse)
library(tigris)
library(sf)

# census_api_key("YOUR KEY HERE", install = TRUE)
options(tigris_class = "sf")
options(tigris_use_cache = TRUE)
```

The examples in this article will illustrate how to build distance profile visualizations for large metropolitan areas in the US state of Texas. To obtain data, the *get_acs()* function from tidycensus is used, which grants access to the five-year ACS APIs. The default API is for 2012–2016 data, the most recent data available at the time of this writing.

```
tx_tracts <- get_acs(geography = "tract",
                     variables = c(hs = "DP02_0061", grad = "DP02_0065"),
                     state = "TX",
                     summary_var = "DP02_0058",
                     geometry = TRUE)

tx_tracts

## Simple feature collection with 10508 features and 7 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -106.6456 ymin: 25.83738 xmax: -93.50829 ymax: 36.50
07
## epsg (SRID):    4269
## proj4string:    +proj=longlat +datum=NAD83 +no_defs
## First 10 features:
##          GEOID                                 NAME variable
## 1  48201343301 Census Tract 3433.01, Harris County, Texas        hs
## 2  48201343301 Census Tract 3433.01, Harris County, Texas      grad
## 3  48201410200     Census Tract 4102, Harris County, Texas        hs
## 4  48201410200     Census Tract 4102, Harris County, Texas      grad
## 5  48201411300     Census Tract 4113, Harris County, Texas        hs
## 6  48201411300     Census Tract 4113, Harris County, Texas      grad
## 7  48201412300     Census Tract 4123, Harris County, Texas        hs
## 8  48201412300     Census Tract 4123, Harris County, Texas      grad
## 9  48201412800     Census Tract 4128, Harris County, Texas        hs
## 10 48201412800     Census Tract 4128, Harris County, Texas      grad
##     estimate moe summary_est summary_moe                      geometry
## 1        904 228        3101         216 MULTIPOLYGON (((-95.08843 2...
## 2         91  50        3101         216 MULTIPOLYGON (((-95.08843 2...
## 3        423 243        4428         669 MULTIPOLYGON (((-95.39842 2...
## 4       1153 267        4428         669 MULTIPOLYGON (((-95.39842 2...
## 5        167  98        2872         329 MULTIPOLYGON (((-95.45659 2...
## 6        851 244        2872         329 MULTIPOLYGON (((-95.45659 2...
## 7        127  97        4899         209 MULTIPOLYGON (((-95.43011 2...
## 8       2333 251        4899         209 MULTIPOLYGON (((-95.43011 2...
## 9        183 101        3464         216 MULTIPOLYGON (((-95.45945 2...
## 10      1660 197        3464         216 MULTIPOLYGON (((-95.45945 2...
```

The function returns an sf tibble where each row represents a unique Census tract-Census variable combination for all tracts in the state of Texas. Two variables from the ACS Data Profile are returned: population age 25 and up where the highest educational attainment is a high school diploma (*DP02_0061*), and population age 25 and up where the highest educational attainment is a graduate degree (*DP02_0065*). The estimates for these variables and their respective margins of error are found in the *estimate* and *moe* columns, respectively. By supplying a variable name to the optional *summary_var* parameter, the function returns *summary_est* and *summary_moe* columns for the total population age 25 and up, which can be used as a denominator in the location quotient calculation. Finally, the *geometry* list-column stores feature geometry for each Census tract.

The next step is to identify the Census core-based statistical area in which each Census tract lies. As tidycensus uses the tigris package to obtain Census feature geometry, the *core_based_statistical_areas()* function in tigris can fetch geometry for CBSAs, which are then matched to the Census tracts in Texas using the *st_join()* function in the sf package. This adds two new columns to the tracts dataset, representing the CBSA ID and name of the CBSA that each tract lies within.

```
metros <- core_based_statistical_areas(cb = TRUE) %>%
  select(metro_id = GEOID, metro_name = NAME)

tx_metro <- st_join(tx_tracts, metros, join = st_within)
```

The first illustration of distance profile visualization will use the Houston metropolitan area, the fifth-largest in the United States by population. As Texas Census tracts are now identified by their corresponding metropolitan areas, the tract dataset can be filtered to only include tracts in the Houston area, as shown in Fig. 1.

```
houston_tracts <- filter(tx_metro, str_detect(metro_name, "Houston"))

# Requires the development version of ggplot2: devtools::install_github("tidy
verse/ggplot2")
ggplot(houston_tracts) + geom_sf()
```

To build the distance profile visualization for Houston, the analyst requires some additional information. This includes the distance between Census tract centroids and Houston city hall, and the location quotients by tract for both graduate degree holders and high school diploma holders. The code below defines a function that will be used to calculate location quotients, and creates a simple feature point object representing the location of Houston City Hall, transformed to the projected coordinate system UTM Zone 15 N, which is appropriate for Houston.
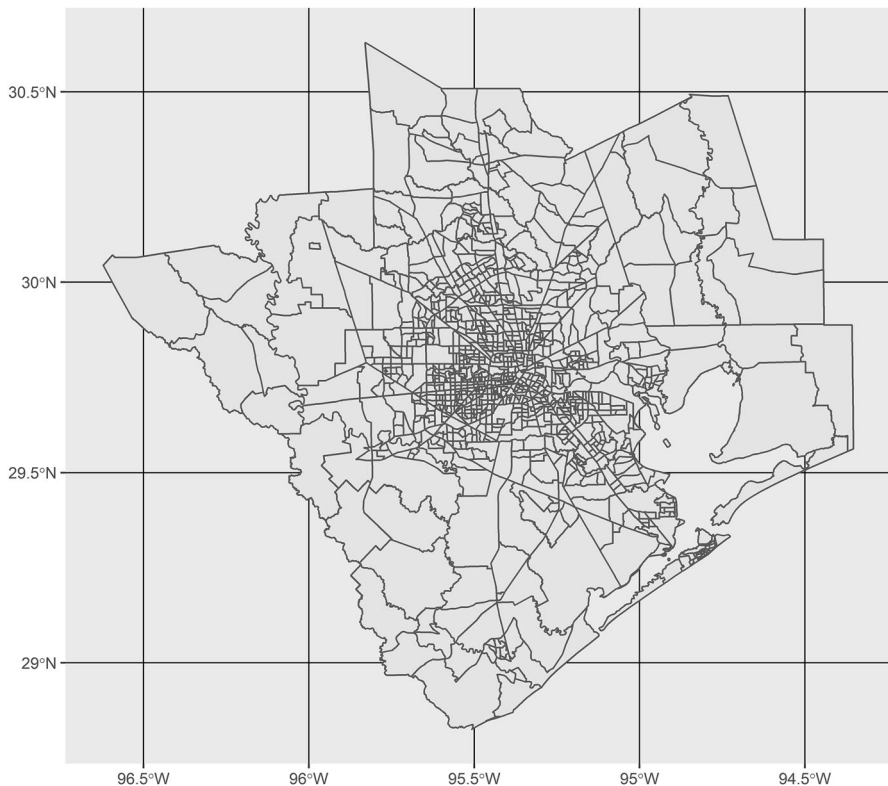
Fig. 1 Census tracts in the Houston, TX metropolitan area

```
lq <- function(group_col, total_col) {
  num <- group_col / total_col
  denom <- sum(group_col, na.rm = TRUE) / sum(total_col, na.rm = TRUE)
  return(num / denom)
}

houston_hall <- c(-95.36976, 29.76044) %>%
  st_point() %>%
  st_sfc(crs = st_crs(houston_tracts)) %>%
  st_transform(26915)
```

Tract distances and location quotients are then calculated within **pipelines**. Code in this article uses the **pipe** operator in R %>% from the **magrittr** package (Milton Bache and Wickham 2014), which is loaded by default with tidyverse. Pipes in R allow analysts to read functions in sequence, separating out steps in the code and avoiding too many complicated nested function calls. The two sequences below could be combined in the same pipeline, but are separated to enhance readability. The first block groups the dataset by ACS variable and calculates location quotients for each variable, to be stored in a column named *lq*; the second transforms the tract

data to the UTM projected coordinate system and stores the distance between tract centroids and Houston City Hall in a column named *dist*.

```
houston_tracts <- houston_tracts %>%
  group_by(variable) %>%
  mutate(lq = lq(estimate, summary_est)) %>%
  ungroup()

houston_tracts <- houston_tracts %>%
  st_transform(26915) %>%
  mutate(dist = as.numeric(
    st_distance(
      st_centroid(.), houston_hall
    ))
  )
```

Once the appropriate columns are calculated in the dataset, distance profile visualization is straightforward. Any of R's plotting libraries would work for this task; **ggplot2** (Wickham 2016) is particularly well-suited as loess smoothing is built into its *geom_smooth()* function. A span parameter of 0.3 is selected in the example, but could be changed by the analyst depending on the desired degree of smoothing. The result is illustrated in Fig. 2.

```
ggplot(houston_tracts, aes(x = dist, y = lq, color = variable)) +
  geom_smooth(method = "loess", span = 0.3)
```

With a few modifications, the plot can be formatted with additional descriptive information about its contents, which is represented in Fig. 3. A dashed line is included for a location quotient value of 1, which represents when Census tract concentration of a population group is the same as that of the metropolitan area as a whole.

```
ggplot(houston_tracts, aes(x = dist / 1000, y = lq, color = variable)) +
  geom_smooth(span = 0.3, method = "loess") +
  geom_hline(yintercept = 1, color = 'black', linetype = 'dashed') +
  theme_minimal() +
  scale_color_brewer(palette = "Dark2",
                     labels = c("Graduate degree", "High school")) +
  labs(x = "Distance from downtown (km)",
       y = "Concentration relative to metropolitan area",
       caption = "Data source: 2012-2016 ACS via the tidycensus R package",
       title = "Houston, TX metropolitan area") +
  theme(legend.title = element_blank(),
        plot.caption = element_text(size = 6))
```

The visualization presents a striking contrast between the geographies of graduate degree holders and high school diploma holders in the Houston metropolitan area. Houston-area residents with graduate degrees are over-represented in neighborhoods near to downtown Houston, and graduate degree holder
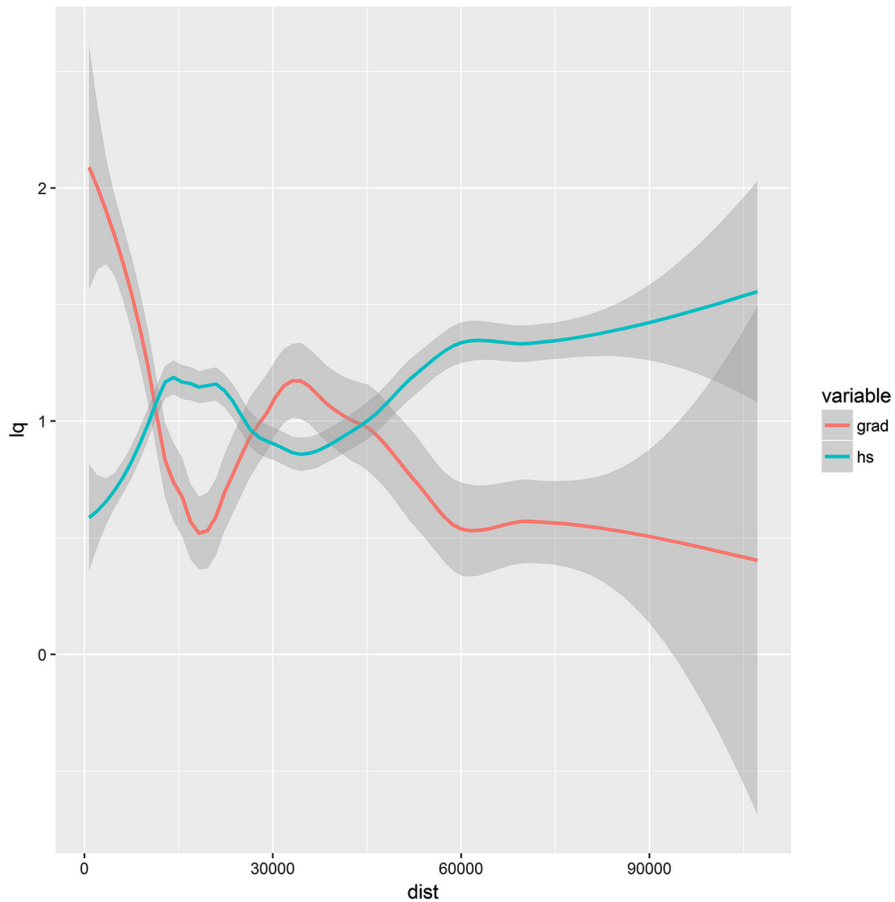
Fig. 2 Distance profile visualization of educational attainment for Houston, TX

concentration then falls with distance from city hall until a distance of about 20 km from downtown. Graduate degree concentration then rises again with a small peak in above-average concentration 35 km from city hall, then falls again toward the rural fringes of the metropolitan areas.

High school diploma holders exhibit a near-opposite profile. They are over-represented on the rural fringes of the Houston metropolitan area and in the urban core between 10 km and 20 km from downtown. Also notable in the visualization is the three points at which the two LOESS curves cross each other. They are suggestive of "zones" of educational attainment within the Houston metropolitan area, in which graduate degree holders are clustered near to downtown and in favored suburbs. Of course, this particular pattern could simply represent Houston-specific characteristics, such as a geographical segmentation of the labor market in which people are living near to jobs commensurate with their level of expertise. As such, distance profile methods are particularly useful as an exploratory method
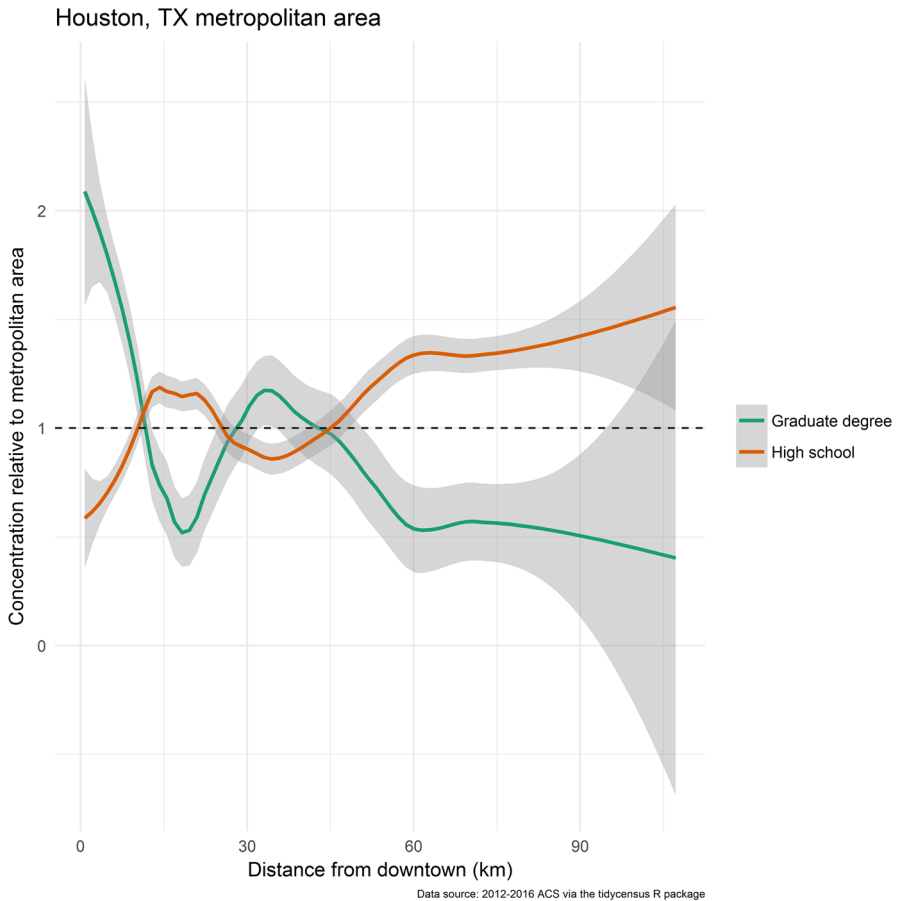
Houston, TX metropolitan area

Fig. 3 Formatted distance profile visualization for Houston

when used in comparative context. The next section will illustrate how to build distance profile visualizations for multiple metropolitan areas.

5 Comparative Distance Profile Visualization

Building distance profile visualizations for multiple metropolitan areas simultaneously presents some additional challenges. For one, the analyst should select an appropriate projected coordinate system for calculating tract centroids for all of the metropolitan areas under study, or alternatively the analyst should iterate through projected coordinate systems appropriate for each metro. Second, some metropolitan areas—like Dallas-Fort Worth—have multiple core cities, meaning that the distance calculation should represent the distance to the *nearest* major city hall to avoid mis-representing Census tracts. Third, in circumstances where metropolitan

areas span multiple states, the analyst will need to obtain tract data for more than one state.

The first example is a comparative visualization of educational distance profiles for the four largest metropolitan areas in Texas: Dallas-Fort Worth, Houston, Austin, and San Antonio. City hall data is obtained from a dataset of the XY coordinates for the city halls of major metropolitan areas in the United States, *city_halls.csv*, which is included with this article. To work with this dataset, the analyst should first read in the CSV file, then specify it as a simple features object, transform to an appropriate projected coordinate system (Texas Centric Albers Equal Area in this example), and filter for the desired metropolitan areas.

```
tx_halls <- read_csv("city_halls.csv") %>%
  st_as_sf(coords = c("X", "Y"), crs = 4269) %>%
  st_transform(3083) %>%
  filter(metro_id %in% c("12420", "19100", "26420", "41700"))

tx_halls

## Simple feature collection with 5 features and 6 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: -470664.8 ymin: 7023527 xmax: 3902501 ymax: 9427499
## epsg (SRID):    3083
## proj4string:    +proj=aea +lat_1=27.5 +lat_2=35 +lat_0=18 +lon_0=-100 +x_0
=1500000 +y_0=6000000 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
## # A tibble: 5 x 7
##                                          metro metro_id
##                                          <chr>    <int>
## 1 Austin-Round Rock-San Marcos, TX Metro Area    12420
## 2  Dallas-Fort Worth-Arlington, TX Metro Area    19100
## 3  Dallas-Fort Worth-Arlington, TX Metro Area    19100
## 4    Houston-Sugar Land-Baytown, TX Metro Area    26420
## 5     San Antonio-New Braunfels, TX Metro Area    41700
## # ... with 5 more variables: address <chr>, city <chr>, state <chr>,
## #   zip <chr>, geometry <simple_feature>
```

There are five rows in the dataset: one each for Austin, Houston, and San Antonio, and two for Dallas-Fort Worth. The approach below then illustrates how to adapt the framework introduced above to account for the two core cities in Dallas-Fort Worth. When *st_distance()* is used to calculate distance between a spatial object and multiple locations, it will return a matrix of the distances to all of the requested locations. As such, for Dallas-Fort Worth, this requires an approach to return the minimum of the two distances calculated.

The approach below uses a map/reduce method to iterate through each metropolitan area, calculate the distance from each tract centroid to its correspond-ing nearest core city hall appropriately, store the result in a list, and then combine the datasets back together for comparative visualization.

```
tx_dist <- map(c("12420", "19100", "26420", "41700"), function(x) {
  m1 <- tx_metro %>%
    filter(metro_id == x) %>%
    st_transform(3083)
  hall <- filter(tx_halls, metro_id == x)
  dist <- st_distance(
    st_centroid(m1), hall
  )
  mindist <- apply(dist, 1, min)
  out <- mutate(m1, dist = mindist)
  return(out)
}) %>%
  reduce(rbind)
```

With the result in hand, the analyst can calculate location quotients by metropolitan area and by variable, and then visualize the distance profiles with ggplot2 as shown above.

```
tx_lq <- tx_dist %>%
  group_by(metro_id, variable) %>%
  mutate(lq = lq(estimate, summary_est)) %>%
  ungroup()

ggplot(tx_lq, aes(x = dist / 1000, y = lq, color = variable)) +
  geom_smooth(span = 0.3, method = "loess") +
  geom_hline(yintercept = 1, color = 'black', linetype = 'dashed') +
  theme_minimal(base_size = 12) +
  facet_wrap(~metro_name) +
  scale_color_brewer(palette = "Dark2",
                     labels = c("Graduate degree", "High school")) +
  labs(x = "Distance from downtown (km)",
       y = "Concentration relative to metropolitan area",
       caption = "Data source: 2012-2016 ACS via the tidycensus R package",
       title = "Educational attainment in Texas metropolitan areas") +
  theme(legend.title = element_blank(),
        legend.position = "bottom",
     plot.caption = element_text(size = 6))
```

The distance profile visualizations for large Texas metropolitan areas in Fig. 4 exhibit several similarities. All have above-average concentrations of graduate degree holders in the suburban rings, although the specific distance associated with this varies based on the size of the metropolitan area. Distinctive differences emerge across the metropolitan areas as well, however. While each metropolitan area shows an uptick in graduate degree holder concentration near to downtown, it is most pronounced in Houston and Austin and less significant in San Antonio. Notably, in Austin graduate degree concentration is at or above average throughout the city center, only falling below average 30 km from downtown, where high school diploma holders rise in concentration. This is not replicated in the other Texas metropolitan areas, who all have above-average concentrations of high school degree holders in select areas within 20 km of their respective downtowns.
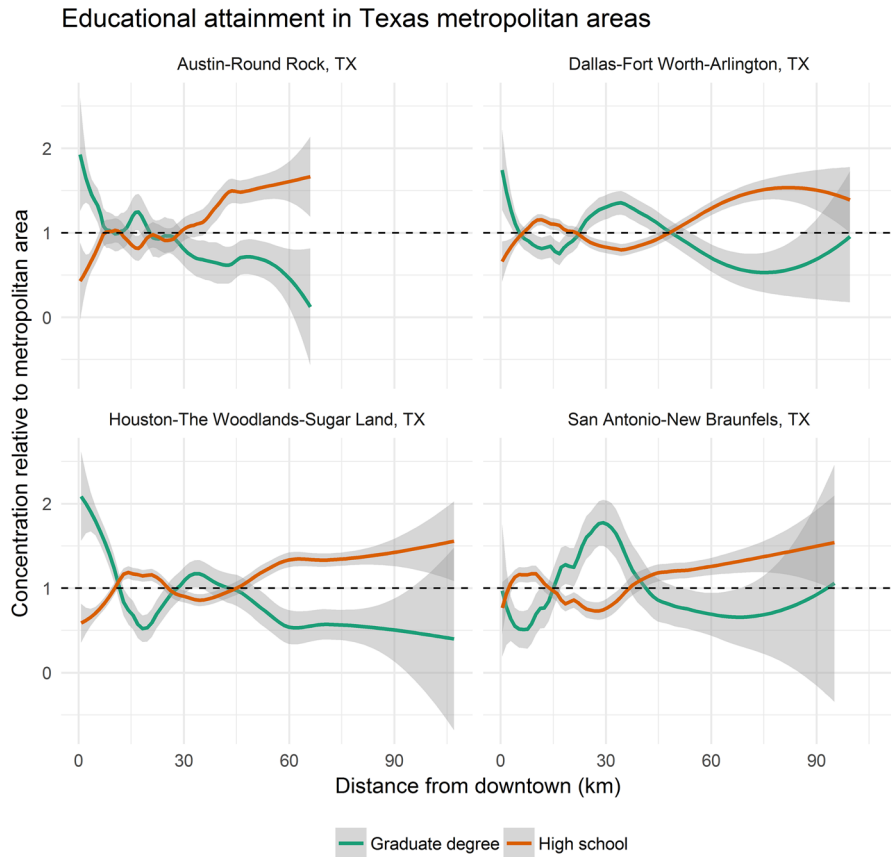
Fig. 4 Distance profile visualizations for major metropolitan areas in Texas

Examining other metropolitan areas only requires slight modifications to the code. The next example demonstrates the code necessary to replicate this for the West Coast metropolitan areas of Seattle, Portland, and San Francisco-Oakland. Like the above example, this workflow will account for the multiple city halls in San Francisco and Oakland, but will also fetch tract data for multiple states, which is necessary for the Portland metropolitan area that includes Census tracts in Oregon and Washington. Figure 5 illustrates the result.

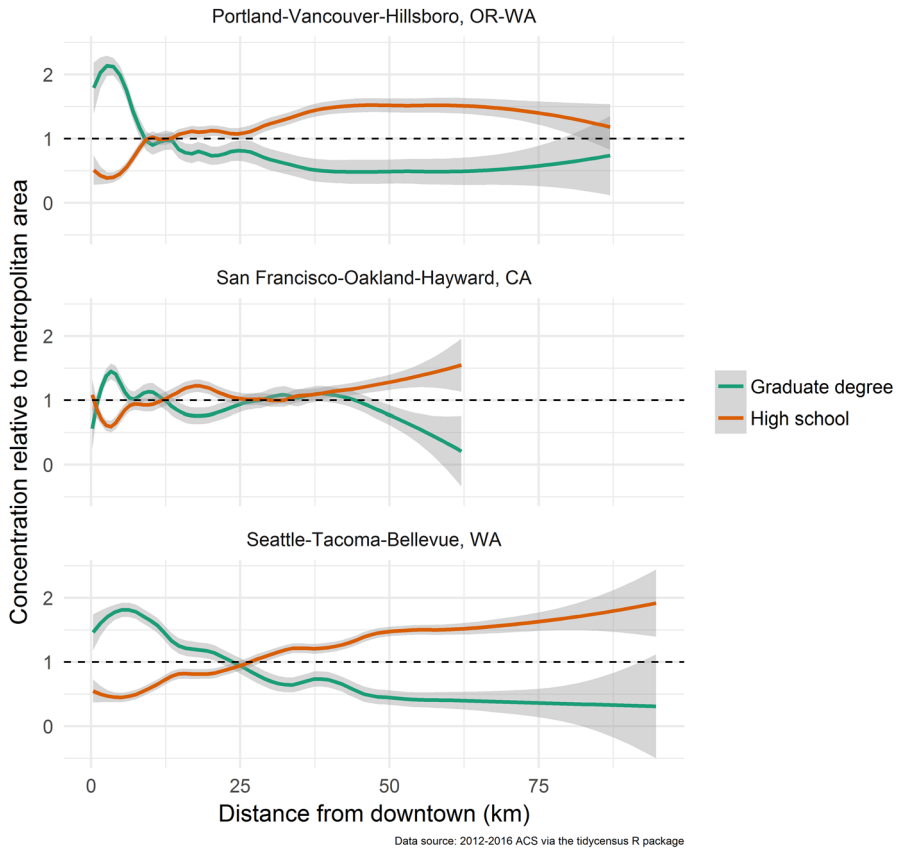**Fig. 5** Distance profile visualizations for three large West Coast metropolitan areas

```r
metros <- core_based_statistical_areas(cb = TRUE) %>%
  select(metro_id = GEOID, metro_name = NAME)

wc_halls <- read_csv("city_halls.csv") %>%
  st_as_sf(coords = c("X", "Y"), crs = 4269) %>%
  st_transform(26910) %>%
  filter(metro_id %in% c("38900", "41860", "42660"))

wc_tracts <- get_acs(geography = "tract",
                     variables = c(hs = "DP02_0061", grad = "DP02_0065"),
                     state = c("CA", "OR", "WA"),
                     summary_var = "DP02_0058",
                     geometry = TRUE) %>%
  st_join(metros, join = st_within) %>%
  st_transform(26910)

wc_lq <- map(c("38900", "41860", "42660"), function(x) {
    m1 <- filter(wc_tracts, metro_id == x)
    hall <- filter(wc_halls, metro_id == x)
    dist <- st_distance(
      st_centroid(m1), hall
    )
    mindist <- apply(dist, 1, min)
    out <- mutate(m1, dist = mindist)
    return(out)
}) %>%
  reduce(rbind) %>%
  group_by(metro_id, variable) %>%
  mutate(lq = lq(estimate, summary_est)) %>%
  ungroup()

ggplot(wc_lq, aes(x = dist / 1000, y = lq, color = variable)) +
  geom_smooth(span = 0.3, method = "loess") +
  geom_hline(yintercept = 1, color = 'black', linetype = 'dashed') +
  theme_minimal(base_size = 13) +
  facet_wrap(~metro_name, ncol = 1) +
  scale_color_brewer(palette = "Dark2",
                     labels = c("Graduate degree", "High school")) +
  labs(x = "Distance from downtown (km)",
       y = "Concentration relative to metropolitan area",
       caption = "Data source: 2012-2016 ACS via the tidycensus R package",
       title = "Educational attainment in three West Coast metropolitan areas
    plot.caption = element_text(size = 6))
```

In Portland and Seattle, graduate degree holders are strongly over-represented in the central city and high school diploma holders over-represented in the suburbs. San Francisco-Oakland exhibits a similar pattern, with the notable exception that high school diploma holders are more prevalent very near to downtown than graduate degree holders. This is likely influenced by the close proximity of San Francisco's downtown to neighborhoods of low educational attainment such as the Tenderloin district and Chinatown, and the inclusion of Oakland City Hall, as Oakland has lower educational attainment overall than San Francisco.

Examination of the Texas and West Coast distance profile visualizations in turn can prompt discussion about similarities and differences between these metropolitan areas. Metropolitan areas in Texas, such as Dallas-Fort Worth, Houston, and Austin, as well as metropolitan areas in the Northwest, such as Seattle and Portland, have high concentrations of graduate degree holders near to the urban core. Neither Seattle nor Portland have prominent graduate degree holder concentrations 30 km from the urban core, however, which are present in Dallas-Fort Worth and Houston. Austin's profile resembles a visual "middle ground" between the Texas and Northwest metropolitan areas, suggesting demographic similarities with Seattle and Portland—all three metros are noted high-technology hubs—and regional similarities with Dallas-Fort Worth and Houston. While such visualizations are exploratory in nature, they do raise these sorts of questions around the influence of labor market structure and regional context on the population geography of US metropolitan areas.

6 Visualizing Distance Profiles Over Time

Distance profile visualization is also a very useful tool for comparing shifts in demographic distributions within metropolitan areas over time. Walker (2018a, b) uses this method to examine changes in the geography of metropolitan racial and ethnic diversity between 1990 and 2010, and Juday (2015b) similarly makes comparisons between 1990 data and more recent ACS data with this method. Comparing distance profiles over time, however, introduces several methodological challenges. For one, Census variable definitions can change over time, so the analyst must take care to ensure that appropriate comparisons are made across Census/ACS years. Additionally, Census *boundaries* shift over time, meaning that such longitudinal analysis is susceptible to modifiable areal unit problem (MAUP) effects, in which the boundaries of the polygons used in a spatial analysis can have a significant influence on the analytical results (Openshaw 1984).

Walker's approach uses Brown University's Longitudinal Tract Database (Logan et al. 2014) for decennial Census data, which interpolates Census tract data since 1970 to 2010 tract boundaries. The LTDB allows analyses to keep tract boundaries consistent over time, potentially limiting MAUP effects. Using interpolated data is not without limitations, however, as historical results are not strictly the decennial Census tabulations but rather allocated estimates. Further, the LTDB only provides a subset of popular Census and ACS variables in its downloadable dataset; to interpolate estimates for other variables they provide scripts for use in Microsoft Access and Stata, which are commercial software packages that are not freely usable.

Given that distance profile visualization is principally a method for data exploration and hypothesis generation, the example below uses the raw 1990 decennial Census data for comparison with data from the 2012–2016 ACS. This is done with the caveat that the results may be influenced by the different tract boundaries employed between the two datasets, and the acknowledgment of differences in sampling design between the 1990 decennial Census and the

American Community Survey. The principal advantage to this approach, however, is that the process can be executed reproducibly without leaving the R environment.

The example below will compare the distance profiles of graduate degree holders in the Houston, TX metropolitan area between the 1990 decennial Census and the 2012–2016 American Community Survey. Decennial Census data in tidycensus can be obtained with the *get_decennial()* function. In 1990, tract data on graduate degree holders age 25 and up is available; however the denominator for the location quotient calculation, all individuals age 25 and up, is split across multiple variables. As such, calculating the location quotient will require an additional call to *get_decennial()*, as illustrated below.

As tract data are only available by county from the 1990 decennial Census API at the time of this writing, the approach below will retrieve data for the specific counties in the Houston metropolitan area as defined by the most recent CBSA definition to keep the comparison as consistent as possible.

```r
houston_area <- filter(metros, metro_id == "26420")

houston_counties <- counties("TX", cb = TRUE) %>%
  st_join(houston_area, join = st_within, left = FALSE) %>%
  pull(COUNTYFP)

houston90a <- get_decennial(geography = "tract",
                            year = 1990,
                            variables = "P0570007",
                            geometry = TRUE,
                            state = "TX",
                            county = houston_counties)

# Get variables for age 25 and up
agevars <- paste0("P01300", 18:31)

houston90b <- get_decennial(geography = "tract",
                            year = 1990,
                            variables = agevars,
                            state = "TX",
                            sumfile = "sf3",
                            county = houston_counties) %>%
  group_by(GEOID) %>%
  summarize(summary_value = sum(value))
```

With the requisite Census data in hand, the datasets can be merged, allowing for a location quotient calculation and then a distance calculation for the distance profile visualization.

```
houston90 <- merge(houston90a, houston90b, by = "GEOID") %>%
  st_transform(26915) %>%
  mutate(year = "1990",
         lq = lq(value, summary_value),
         dist = as.numeric(
           st_distance(
             st_centroid(.), houston_hall
             )
           )
         ) %>%
  select(GEOID, year, lq, dist)
```

To build the comparative visualization, the 1990 dataset is combined with the 2012–2016 ACS dataset. Side-by-side faceted charts are used given that the 1990 and 2012–2016 datasets reflect both different tract boundaries and different sampling designs.

```
houston_combined <- houston_tracts %>%
  filter(variable == "grad") %>%
  mutate(year = "2012-2016") %>%
  select(GEOID, year, lq, dist) %>%
  rbind(houston90)

ggplot(houston_combined, aes(x = dist / 1000, y = lq)) +
  geom_smooth(span = 0.3, method = "loess", color = "blue") +
  geom_hline(yintercept = 1, color = 'black', linetype = 'dashed') +
  theme_minimal() +
  facet_wrap(~year) +
  labs(x = "Distance from downtown (km)",
       y = "Concentration relative to metropolitan area",
       caption = "Data source: the tidycensus R package",
       title = "Graduate degree holders in Houston, TX") +
  theme(plot.caption = element_text(size = 6))
```

The charts for each Census/ACS sample in Fig. 6 exhibit some similarities. Graduate degrees are rarer on the metropolitan fringe in both plots, and over-represented in both suburbs 30 km from the urban core and in the central city, reflecting the stable affluent areas in Houston to the west of downtown. The most distinctive change, however, represents the areas very near to downtown. Whereas graduate degree holders in 1990 were strongly under-represented within 5 km of city hall, they are strongly over-represented in 2012–2016. This shift reflects a process of downtown gentrification that has characterized Houston's metropolitan demography over the past two decades (Holeywell 2013). Additional comparative visualization could reveal whether this trend is replicated across metropolitan areas in the US, which could be accomplished using the methods presented in this article.

7 Exploring Demographic Distance Profiles Geographically

Distance profile visualization offers a useful exploratory framework for illustrating demographic shifts relative to the urban core in metropolitan areas. However, the LOESS curves may fail to pick up notable outliers, such as demographic clusters of
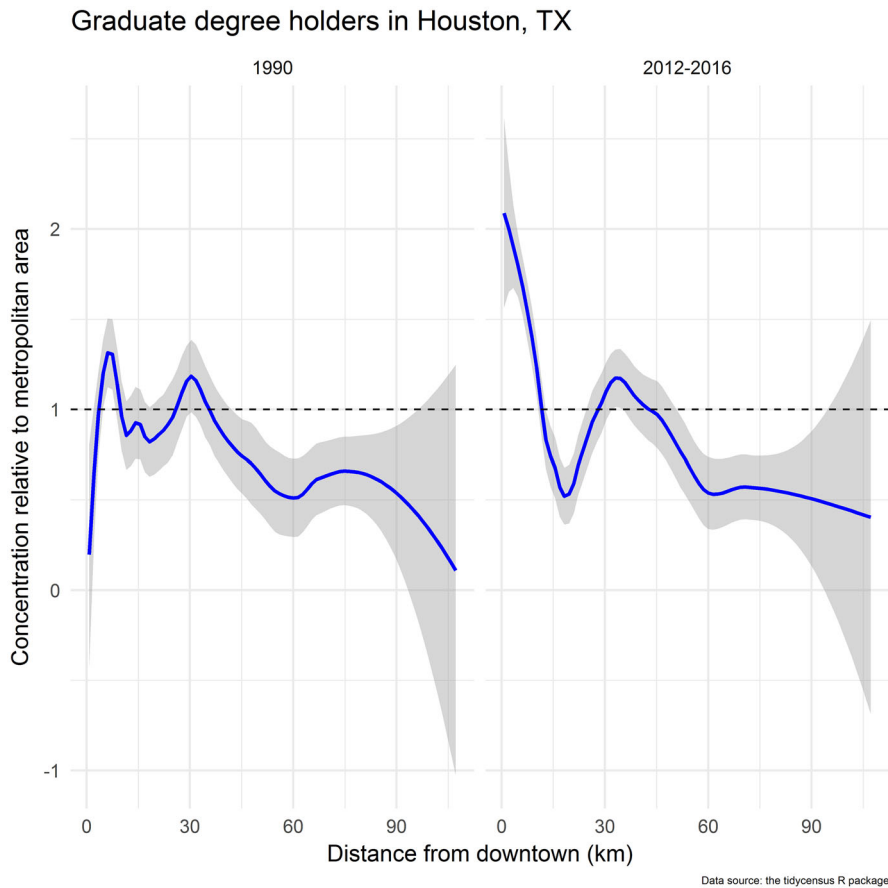
Graduate degree holders in Houston, TX

Fig. 6 Comparative distance profile visualization for Houston, TX between 1990 and 2012–2016

highly educated individuals living distant from urban cores. Tools within the R environment allow for additional exploration of these issues through mapping and linked brushing.

An example of this type of approach is found in Walker (2018a, b), which analyzes neighborhood diversity through both distance profile visualization, which accounts for general trends, and exploratory spatial data analysis, which identifies spatial clusters of high- and low-diversity neighborhoods. R further allows for the development of interactive tools to explore distance profile visualizations in more detail. To accomplish this, Walker developed an interactive dashboard using R's Shiny framework to accompany the published article. The dashboard allows users to select Census tract points on a distance profile visualization to highlight and zoom to the corresponding Census tracts on a linked interactive map. In turn, the dashboard can provide additional context to readers of the research and researchers who might want to explore the article's content in more detail (Misra 2016).

To first explore a potential lack of fit in the LOESS smoother, an analyst might consider mapping the residuals of the local regression model by Census tract. As tidycensus spatial objects already include feature geometry, making residual maps is straightforward with the *geom_sf()* function available in ggplot2. The first example returns to the *houston_tracts* object used to introduce the concept of distance profile visualization earlier in the paper, but filters it to only return rows representing graduate degree holders. While the wrapper *geom_smooth()* in ggplot2 was used to create the LOESS visualization in the earlier example, the R function *loess()* can store the local regression model in an object, named *l1* below.

```
houston_grad <- filter(houston_tracts, !is.na(lq), variable == "grad")

l1 <- loess(lq ~ dist, data = houston_grad, span = 0.3)
```

The vector *l1$residuals* stores the model residuals for Houston's distance profile, which in turn can be added to the *houston_grad* object as a column and then mapped (Fig. 7).
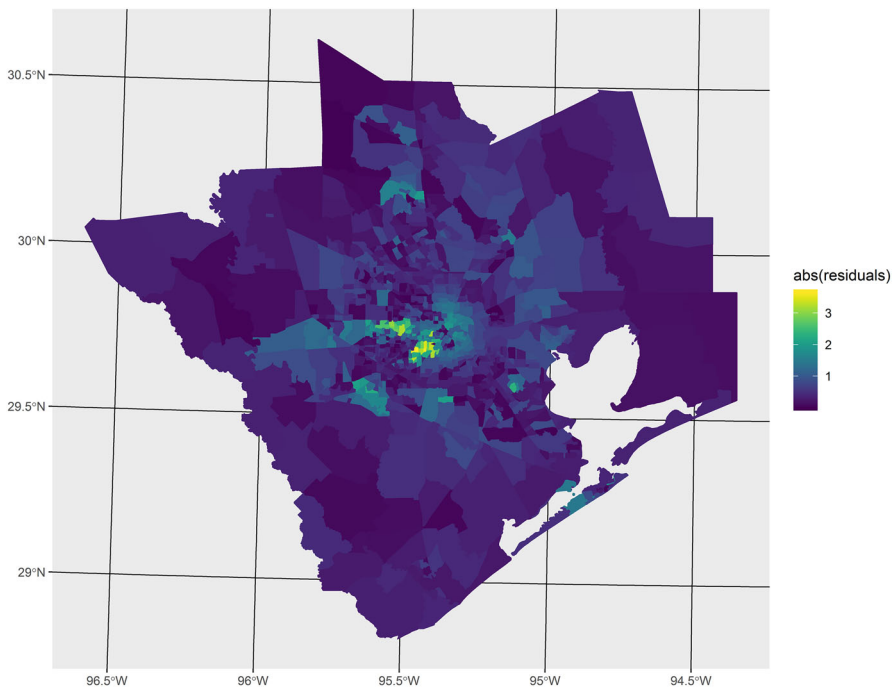

Fig. 7 Map of absolute values of residuals from a LOESS model used in distance profile visualization of graduate degree-holder concentration in the Houston, TX metropolitan area

```
library(viridis)

houston_grad$residuals <- abs(l1$residuals)

ggplot(houston_grad, aes(fill = residuals, color = residuals)) +
  geom_sf() +
  scale_color_viridis() +
  scale_fill_viridis() +
  labs(fill = "abs(residuals)",
       color = "abs(residuals)")
```

The map shows the absolute values of residuals from the model, illustrating areas where the model fits poorly in bright yellow. We note pockets within the center of the graphic—representing parts of central Houston—with the most significant lack of fit. These neighborhoods may represent neighborhoods of very high—or very low—educational attainment in the urban core that get smoothed over by the LOESS model given the extreme levels of educational inequality that can exist within urban cores.

Residual exploration can be augmented further by tools in R for linked brushing of map and scatterplot data. Walker (2016a, b) details the utility of the **plotly** R package for interactive demographic data visualization (Sievert et al. 2017), allowing for rapid exploration of demographic data by analysts. plotly can further be used to establish linkages between different charts through the R **crosstalk** package (Cheng 2017).

To accomplish this, the analyst initializes a shared data object linked by tract GEOID from the Houston tract data, then generates a distance profile visualization and residual map as illustrated in this paper. The key is that the *ggplotly()* function in plotly can be used to convert the static ggplot2 graphics to interactive plotly
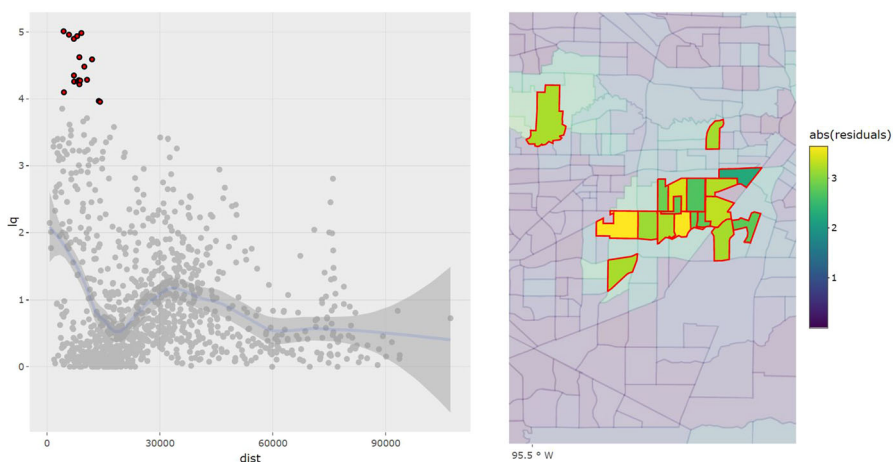

Fig. 8 Linked brushing of points on distance profile visualization and map of absolute values of residuals using plotly and crosstalk

visualizations, which will then respond to user brushing by calling the *highlight()* function in plotly (Fig. 8).

```r
library(plotly)
library(crosstalk)

houston_shared <- SharedData$new(houston_grad, key = ~GEOID)

dp <- ggplot(houston_shared, aes(x = dist, y = lq)) +
  geom_point() +
  geom_smooth(method = "loess", span = 0.3)

map <- ggplot(houston_shared, aes(fill = residuals, color = residuals)) +
  geom_sf() +
  scale_fill_viridis() +
  scale_color_viridis() +
  labs(fill = "abs(residuals)",
       color = "abs(residuals)")

dp_plotly <- ggplotly(dp, height = 600) %>%
  layout(dragmode = "lasso") %>%
  highlight("plotly_selected", color = "red")

map_plotly <- ggplotly(map, height = 600) %>%
  highlight("plotly_selected", color = "red")

bscols(dp_plotly, map_plotly)
```

By brushing the points on the scatterplot/distance profile visualization to the left (which are highlighted in red), the analyst can in turn highlight the corresponding Census tracts in the map to the right. In this instance the highlighted tracts represent neighborhoods in west Houston near Rice University with very high levels of educational attainment. In turn, the exploratory framework detailed in this paper can assist the analyst with visual representation of metropolitan-level trends, but also can facilitate exploration of localized patterns and/or outliers, as evidenced in this section.

8 Conclusion

This article has covered a reproducible framework using the R software environment for visualizing demographic distance profiles from decennial Census and American Community Survey data, both across metropolitan areas and over time. The concept of a demographic distance profile has a long history in urban studies research, and distance profile visualization is an effective way to communicate graphically about distance profiles. The framework outlined in this paper can help researchers and practitioners use distance profile visualization in their own projects, given that the code can be executed entirely within the free, open-source, and cross-platform R programming language.

Distance profile visualization is primarily an exploratory method for helping understand the geographic distribution of a particular population group in a metropolitan area. As mentioned earlier in the paper, care should be taken to acknowledge MAUP effects, especially when analyzing changes in distance profiles

over time. Other MAUP effects may be introduced by the use of the Census tract in this article, which can exhibit internal demographic heterogeneity. Visualization results could in turn be checked against the same results for block groups, which are also available in the tidycensus package.

The distance profile visualization method is ideally used as part of a broader project to aid in visual communication or hypothesis generation. The smoothed relationship represented in distance profile visualizations may in fact reflect the influence of one or more latent variables, and should not be interpreted strictly as a measure of group residential "preference" without additional information. For example, one could envision a simplistic interpretation of Seattle's distance profile as that graduate degree holders crave the cultural opportunities afforded in central Seattle, whereas high school diploma holders desire the cultural sterility of the suburbs. Such an explanation ignores other potential factors, such as that graduate degree holders may be more likely to be able to afford the expensive rents in Seattle's urban core as opposed to high school graduates. Generalized additive models (GAMs) allow a potential path forward here, as they allow the estimation of a response variable as a function of one or more linear or smooth functions of predictors (Wood 2017). Partial regression plots of GAM results can then be used to assess whether distance profiles remain robust when holding other covariates constant.

Even if the visualized distance profile is in fact a function of other latent covariates, the distance profile visualization has proven useful as it has helped generate hypotheses for modeling of these relationships. As this article illustrates, a process that once required multiple software packages—possibly including dedicated GIS software—and cumbersome external data downloads can now be executed entirely within the R computing environment so long as the city hall locations are known. Executing the framework does require expertise on behalf of the analyst, particularly in the areas of Census/ACS variable selection, knowledge of coordinate systems, and R programming. For those inclined analysts, this framework can assist with the reproducible application of distance profile visualization in their own urban demographic projects.

Compliance with Ethical Standards

Conflict of interest The author declares no conflicts of interest regarding this manuscript.

References

Alba, R. D., & Logan, J. R. (1991). Variations on two themes: Racial and Ethnic Patterns in the Attainment of Suburban Residence. *Demography, 28*(3), 431–453.

Albouy, D., & Lue, B. (2015). Driving to opportunity: Local rents, wages, commuting, and sub-metropolitan quality of life. *Journal of Urban Economics, 89,* 74–92.

Alonso, W. (1964). *Location and land use*. Cambridge, MA: Harvard University Press.

Alperovich, G., & Deutsch, J. (1992). Population density gradients and urbanisation measurement. *Urban Studies, 29*(8), 1323–1328.

Berry, B. J. L., Simmons, J. W., & Tennant, R. J. (1963). Urban population densities: Structure and change. *Geographical Review, 53*(3), 389–405.

Brown, E., & Shapiro, T. R. (2015). Cities are becoming more affluent while poverty is rising in inner suburbs—And that has implications for schools. The Washington Post, February 26. https://www.washingtonpost.com/news/local/wp/2015/02/26/cities-are-becoming-more-affluent-while-poverty-is-rising-in-inner-suburbs-and-that-has-implications-for-schools/?utm_term=.d5ebee53cce2. Accessed 27 Feb 2018.

Cheng, J. (2017). Crosstalk: Inter-widget interactivity for HTML widgets. R package version 1.0.1. https://rstudio.github.io/crosstalk/.

Clark, C. (1951). Urban population densities. *Journal of the Royal Statistical Society, Series A (General)*, *114*(4), 490–496.

Eberts, R. W. (1981). An empirical investigation of intraurban wage gradients. *Journal of Urban Economics, 10*(1), 50–60.

Edsall, T. B. (2015). The gentrification effect. *The New York Times*, February 25. https://www.nytimes.com/2015/02/25/opinion/the-gentrification-effect.html?_r=0.

Estiri, H., & Krause, A. (2016). A Cohort Location Model of household sorting in US metropolitan regions. *Urban Studies*. https://doi.org/10.1177/0042098016668783.

Estiri, H., Krause, A., & Heris, M. P. (2015). "Phasic" metropolitan settlers: A phase-based model for the distribution of households in US metropolitan regions. *Urban Geography, 36*(5), 777–794.

Gong, Y., Beolhouwer, P., & de Haan, J. (2016). Interurban house price gradient: Effect of urban hierarchy distance on house prices. *Urban Studies, 53*(15), 3317–3335.

Hackworth, J. (2005). Emergent urban forms, or emergent post-modernisms? A comparison of large U.S. metropolitan areas. *Urban Geography, 26*(6), 484–519.

Holeywell, R. (2013). Houston: The surprising contender in America's urban revival. *Governing*, October. http://www.governing.com/topics/urban/gov-houston-urban-revival.html. Accessed 11 Feb 2018.

Juday, L. J. (2015a). The changing shape of American cities. Demographics Research Group, Weldon Cooper Center for Public Service, University of Virginia. http://www.coopercenter.org/sites/default/files/node/13/ChangingShape-AmericanCities_UVACooperCenter_March2015.pdf. Accessed 11 Feb 2018.

Juday, L. J. (2015b). The changing shape of America's metro areas. http://statchatva.org/changing-shape-of-american-cities/. Accessed 11 Feb 2018.

Lahiri, K., Lankford, R. H., & Numrich, R. P. (1989). The estimation and interpretation of urban density gradients. *Journal of Business and Economic Statistics, 7*(2), 227–235.

Logan, J. R., Xu, Z., & Stults, B. J. (2014). Interpolating U.S. decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer, 66*(3), 412–420.

McMillen, D. P., & Singell, L. D. (1992). Work location, residence location, and the intraurban wage gradient. *Journal of Urban Economics, 32*(2), 195–213.

Milton Bache, S., & Wickham, H. (2014). magrittr: A forward-pipe operator for R. R Package version 1.5. https://CRAN.R-project.org/package=magrittr.

Misra, T. (2016). Where are the most diverse neighborhoods? *CityLab* https://www.citylab.com/equity/2016/05/a-new-visual-tool-explores-the-changing-geography-of-diversity-in-us-metros-app/481218/. Accessed 11 Feb 2018.

Muth, R. F. (1969). *Cities and housing: The spatial pattern of urban residential land use*. Chicago: University of Chicago Press.

Office of Management and Budget. (2017). OMB Bulletin No. 17-01: Revised delineations of metropolitan statistical areas, micropolitan statistical areas, and combined statistical areas, and

guidance on uses of the delineations of these areas. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/bulletins/2017/b-17-01.pdf. Accessed 11 Feb 2018.

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich, UK: Geobooks.

Pebesma, E. (2017). sf: Simple features for R. R package version 0.5-0. https://github.com/edzer/sfr.

Sander, W. (2006). Educational attainment and residential location. *Education and Urban Society, 38*(3), 307–326.

Schnore, L. F. (1963). The socio-economic status of cities and suburbs. *American Sociological Review, 28*(1), 76–85.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2017). plotly: Create interactive web graphics via 'plotly.js'. R package version 4.7.1. https://CRAN.R-project.org/package=plotly.

Sparks, C. (2014). Measuring residential segregation using R: So long to factfinder. *Spatial Demography, 2*(1), 72–78.

Taggart, L. J. (1971). Another look at the burgess hypothesis: Time as an important variable. *American Journal of Sociology, 76*(6), 1084–1093.

United States Census Bureau (2017). American Community Survey: Data profiles. https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2015/. Accessed June 11, 2017.

Walker, K. E. (2016a). tigris: An R package to access and work with geographic data from the US Census Bureau. *The R Journal, 8*(2), 231–242.

Walker, K. E. (2016b). Tools for interactive visualization of global demographic indicators in R. *Spatial Demography, 4*(3), 207–220.

Walker, K. E. (2017). The shifting destinations of metropolitan migrants in the U.S., 2005–2011. *Growth and Change*. https://doi.org/10.1111/grow.12187.

Walker, K. E. (2018a). Locating neighbourhood diversity in the American metropolis. *Urban Studies, 55*(1), 116–132.

Walker, K. E. (2018b). tidycensus: Load US Census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames. R package version 0.4.1. https://walkerke.github.io/tidycensus.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1–23.

Wickham, H. (2016). *ggplot2: Elegant graphics of data analysis*. New York: Springer.

Wickham, H., & Grolemund, G. (2017). *R for data science*. Sepastopol, CA: O'Reilly Media.

Wilson, S. G., Plane, D. A., Mackun, P. J., Fischetti, T. R., & Goworowska, J. (2012). *Patterns of metropolitan and micropolitan population change: 2000 to 2010*. 2010 Census Special Reports, US Census Bureau. https://www.census.gov/prod/cen2010/reports/c2010sr-01.pdf. Accessed 11 Feb 2018.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Boca Raton, FL: CRC Press.