Parcel-level metrics for evaluating housing sites

Carole Voulgaris and Elizabeth Christoforetti

2022-06-17

Contents

Chapter 1

About

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

1.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

1.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

1.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 2

Introduction

Motivation for the project.

Chapter 3

Related work

Talk about various area-level metrics

- Sprawl index
- Neighborhood typology

Talk about VMT site work

Chapter 4

Methodology

4.0.1 Variables

4.0.1.1 Categories that would be useful (things to predict?)

Owner-occupied? Investor-owned? Vacant? Demolition in the past year (no construction since) Construction in the past year

4.0.1.2 factor analysis Variables

The variables made it into the initial factor analysis were:

- Accessibility
  - Distance to transit (use number of transit stops within 1/2 mile walkshed)
  - Share of old/new homes (use average age of homes within 1/2 mile walkshed)
  - Transit frequency (use transit stops per hour within 1/2 mile walkshed)
- Affordability
  - Average Condition of homes in half-mile walkshed
  - Median rent of block-groups with centroids within 1/2 mile walkshed.
  - Median income of block groups with centroids within 1/2 mile walkshed.
  - Median ownership cost of block groups with centroids within 1/2 mile walkshed.

- Close
  - –
- Diverse buildings
  - Entropy of housing types (apartment, townhomes, etc) within 1/2 mile walkshed
- Other
  - Standard deviation of building age within 1/2 mile walkshed

4.1 Data

We obtained data on property addresses, land uses, assessed values (for both land and buildings), and the dates and prices of as many as the three most-recent sales from
Allegheny County Office of Property Assessments [2022], which includes information on 582,116 properties in Allegheny County.

We also obtained latitude and longitude coordinates for each property from a geocoder file provided by Western Pennsylvania Regional Data Center [2021]. Over 99.5 percent of properties included in the assessment dataset are included in the geocoder file. Properties without geocoded locations are excluded from our analysis.

Potential development sites were identified as those

1. classified as "residential" (indicating residential properties with one to four housing units) or "commercial" (which includes mixed-use developments and residential properties with more than four housing units), and
2. with a land use description in one of 59 possible categories[1]. The most common of these are listed Table 4.1.[2].

[1] One site (3008 Phillip Dr in Clairton) is missing a land use description in the assessment data. We checked this address on Zillow to determine that this is a single-family home and classified it as such in our data.

[2] The land use descriptions that were classified as potential development sites but are not listed in Table 4.1, which combine to represent less than one percent of all sites are "RIGHTOF WAY - RESIDENTIAL", "CONDOMINIUM UNIT", "DWG USED AS OFFICE", "APART:20-39 UNITS", "CONDO GARAGE UNITS", "COMMON AREA", "CONDO DEVELOPMENTAL LAND", "CONDEMNED/BOARDED-UP", "CONDOMINIUM OFFICE BUILDING", "INDEPENDENT LIVING (SENIORS)", "DWG USED AS RETAIL", "OTHER COMMERCIAL", "MOBILE HOMES/TRAILER PKS", "RIGHT OF WAY - COMMERCIAL", "GROUP HOME", "TOTAL/MAJOR FIRE DAMAGE - COMM", "OTHER COMMERCIAL HOUSING", "TOTAL/MAJOR FIRE DAMAGE", "COMM APRTM CONDOS 5-19 UNITS", "MUNICIPAL URBAN RENEWAL", "COM-

Table 4.1: Most common land uses categorized as potential sites

USEDESC	Number of potential sites	Percent of potential sites	Cumulative percent of potential sites
SINGLE FAMILY	370,513	73.2	73.2
VACANT LAND	62,672	12.4	85.5
TWO FAMILY	17,293	3.4	89.0
TOWNHOUSE	14,670	2.9	91.8
ROWHOUSE	11,082	2.2	94.0
VACANT COMMERCIAL LAND	5,817	1.1	95.2
THREE FAMILY	3,968	0.8	96.0
RES AUX BUILDING (NO HOUSE)	3,601	0.7	96.7
RETL/APT'S OVER	3,354	0.7	97.3
COMM AUX BUILDING	2,825	0.6	97.9
APART: 5-19 UNITS	2,771	0.5	98.4
FOUR FAMILY	2,058	0.4	98.9
BUILDERS LOT	1,230	0.2	99.1
PARKING GARAGE/LOTS	891	0.2	99.3
OFFICE/APARTMENTS OVER	854	0.2	99.4
MOBILE HOME	666	0.1	99.6
APART:40+ UNITS	529	0.1	99.7
DWG USED AS OFFICE	440	0.1	99.8
APART:20-39 UNITS	400	0.1	99.8
CONDEMNED/BOARDED-UP	132	0.0	99.9

Potential building sites were further filtered to exclude those with missing data on the most recent sale (about one percent of all sites).[3] for a total of potential sites.

The focus of this analysis is on potential development sites rather than on properties. Some properties in the assessor dataset are condominums where multiple properties share a single parcel of land. We aggregated these to the site level by identifying all properties with an assessed building value greater than zero, a land value of zero, and a land use description that did not indicate the land was vacant. If multiple such properties share an address, we classified all properties at that address as a condominium and aggregated them to the parcel level. This led to a final sample of 518,032 sites.

4.1.1 Tax assessment data

Three variables (total assessed fair market value, assessed fair market value of the building, and lot area) were taken directly from the county tax assessment data for use in our analysis. We also included the most recent listed sales price, adjusted for inflation.

To aggregate properties identified as condominiums to the site level, we summed the total values for lot area, assessed land value, assessed building value, and inflation-adjusted sale price. We log-transformed these four variables prior to including them in our analysis. Their distributions are shown in Figure 4.1.

MERCIAL LAND", "CAMPGROUNDS", "COMMON AREA OR GREENBELT", "CHARITABLE EXEMPTION/HOS/HOMES", "INCOME PRODUCING PARKING LOT", "DWG APT CONVERSION", ">10 ACRES VACANT", "MINOR FIRE DAMAGE", "COMM APRTM CONDOS 20-39 UNITS", "COMMERCIAL/UTILITY", "H.O.A RECREATIONS AREA", "COMM APRTM CONDOS 40+ UNITS", "MINOR FIRE DAMAGE - COMM", "OTHER", "OTHER RESIDENTIAL STRUCTURE", "OWNED BY METRO HOUSING AU", "RESIDENTIAL VACANT LAND", "HUD PROJ #221", and "VACANT LAND 0-9 ACRES"

[3]Four sites had sales prices listed that were unreasonably high. 3039 Liberty Avenue in Pittsburgh is listed as having sold for $511,945,000 on August 30, 2021. Zillow lists this property as having sold on that date for $511,945 (https://www.zillow.com/homedetails/3039-W-Liberty-Ave-Pittsburgh-PA-15216/2070262638_zpid/, accessed 5/4/2022), so the value was corrected for what appears to have been a typo. 220 Hyeholde Dr in Coraopolis is listed as having sold for $28,100,000 in 1967. This may also be a typo, and it also does not seem to be the most recent sale. Zillow lists this home as having sold for $350,000 in 2004 (https://www.zillow.com/homes/220-hyeholde-dr,-Coraopolis,-PA_rb/11552817_zpid/, accessed 5/4/2022), so the data was corrected to add that as the most recent sale. Two other sites were identified as having unreasonably high sales values: 1339 Arlington Avenue in Pittsburgh is a three-bedroom single-family home that is listed as having sold for $57,010,813 in 1976 and a 0.06-acre vacant lot with tax ID 0165G00270000000 is listed as having sold for $24,920,232 in 1936. The sales data for these sites were treated as missing.

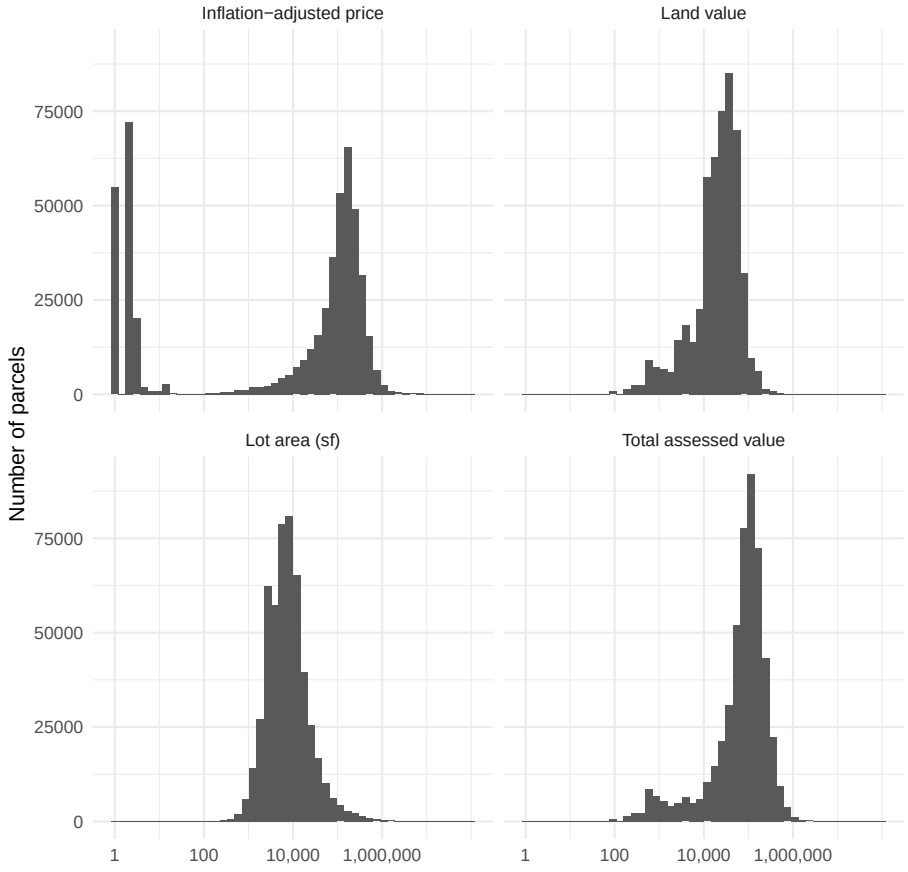

Figure 4.1: Distribution of variables from tax assessor database

4.1.2 Accessibilty data

Accessibilty was calculated from each of the 518,032 sites in our sample to each of several location types described below.

4.1.2.1 Destination parcels

We used land use codes from the county assessor parcel data to identify *destination parcels* that residents might value access to. The most common land use codes of identified destination parcels are listed in Table 4.2.

4.1.2.2 Job locations

We identified *job locations* based on data from a Longitudinal Employer-Household Dynamics (LEHD) dataset published by the United States Census Bureau [United States Census Bureau, 2021]. The LEHD dataset provides the total number of jobs in each census block in the United States, based on employment tax records. The location of each job was defined as the centroid of the block in which it was located. We downloaded job location data for Pennsylvania and filtered it to include locations in the Pittsburgh metropolitan area (Allegheny, Armstrong, Beaver, Butler, Fayette, Washington, and Westmoreland counties).

In addition to calculating the accessibility to jobs of all categories, we also calculated accessibility to several subsets of jobs. We disaggregated jobs by earnings, reasoning that the usefulness of a job might vary depending on how well it matches a workers skills or wage expectations. *High-paying job locations* are a subset of job locations where the worker earns more than $3333 per month. *Low-paying job locations* are those where the worker earns $1250 per month or less.

We also disaggregated jobs based on employment industry, based on the North American Industry Classification System (NAICS), reasoning that the presence of jobs particular industries might represent a shopping or recreation destination. *Retail job locations* are a subset of job locations in NAICS sector 44-45 (retail trade); *Entertainment job locations* are those in NAICS sector 71 (arts, entertainment, and recreation); and *Hospitality job locations* are those in NAICS sector 72 (accommodation and food services).

Finally, we identified three location types that correspond with common non-work trips: schools, grocery stores, and parks. *Grocery store locations* were identified as vendors participating in the Supplemental Nutrition Program for Women, Infants, and Children (WIC). WIC vendor locations and *school locations* were obtained from the Allegheny County GIS portal [Allegheny County Office of Information Technology, 2018, 2020]. *Park locations* were taken from the Pennsylvania Geospatial Data Clearinghouse [Pennsylvania Department of

Table 4.2: Land uses identified as potential destinations

USEDESC	Number of identified destinations	Percent of identified destinations	Cumulative percent of identified destinations
MUNICIPAL GOVERNMENT	10,376	29.88	29.88
CHURCHES, PUBLIC WORSHIP	1,946	5.60	35.49
COMMERCIAL GARAGE	1,735	5.00	40.48
OFFICE - 1-2 STORIES	1,649	4.75	45.23
SMALL DETACHED RET(UNDER 10000)	1,646	4.74	49.97
OFFICE/WAREHOUSE	1,386	3.99	53.96
COUNTY GOVERNMENT	1,287	3.71	57.67
WAREHOUSE	1,252	3.61	61.27
OWNED BY BOARD OF EDUCATION	1,086	3.13	64.40
TOWNSHIP GOVERNMENT	855	2.46	66.86
LIVESTOCK O/T D & P-CAUV	805	2.32	69.18
LIGHT MANUFACTURING	799	2.30	71.48
PUBLIC PARK	710	2.04	73.53
RESTAURANT, CAFET AND/OR BAR	697	2.01	75.54
GENERAL FARM	607	1.75	77.28
OWNED BY COLLEGE/UNIV/ACADEMY	458	1.32	78.60
MEDICAL CLINICS/OFFICES	445	1.28	79.88
RETL/OFF OVER	442	1.27	81.16
OFFICE-ELEVATOR -3 + STORIES	412	1.19	82.34
LODGE HALL/AMUSEMENT PARK	386	1.11	83.46
AUTO SALES & SERVICE	363	1.05	84.50
RETL/STOR OVER	344	0.99	85.49

Conservation and Natural Resources, 2015]. Park locations were downloaded for Pennsylvania and filtered to Allegheny county.

We used the r5r package in the R programming language [Pereira et al., 2021] to calculate accessibility each destination type described above, for each of four transportation modes (walking, cycling, driving, and transit). The r5r package calculates accessibility as the weighted total number of destinations reachable by a given mode, where destinations are weighted according to a decay function, such that destinations that can be reached within less time are assigned greater weight. We used a logistic decay function, as illustrated in 4.2. For motorized modes, the decay function had a mean (inflection) of 40 minutes and a standard deviation of 10 minutes. For non-motorized modes, the decay function had a mean of 20 minutes and a standard deviation of 5 minutes.

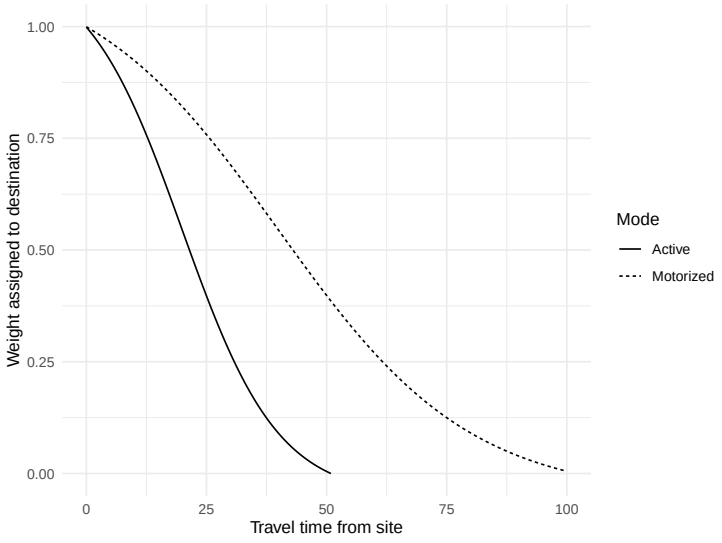

Figure 4.2: Decay functions for accessibility calculations

Calculating accessibility metrics for a combination of four transportation modes and ten destination types yields 40 different accessibility variables. 4.2 illustrates the distributions of each of these variables.

4.1.3 Disamenity proximity

We categorized several land uses in the county assessor data as disamenities. The land use codes we used to identify disamenities are listed in **??**[4].

[4]289 properties related to coal mining (with land use descriptions of either "COAL RIGHTS, WORKING INTERESTS" or "COAL LAND, SURFACE RIGHTS") are co-located and are treated as a single site.

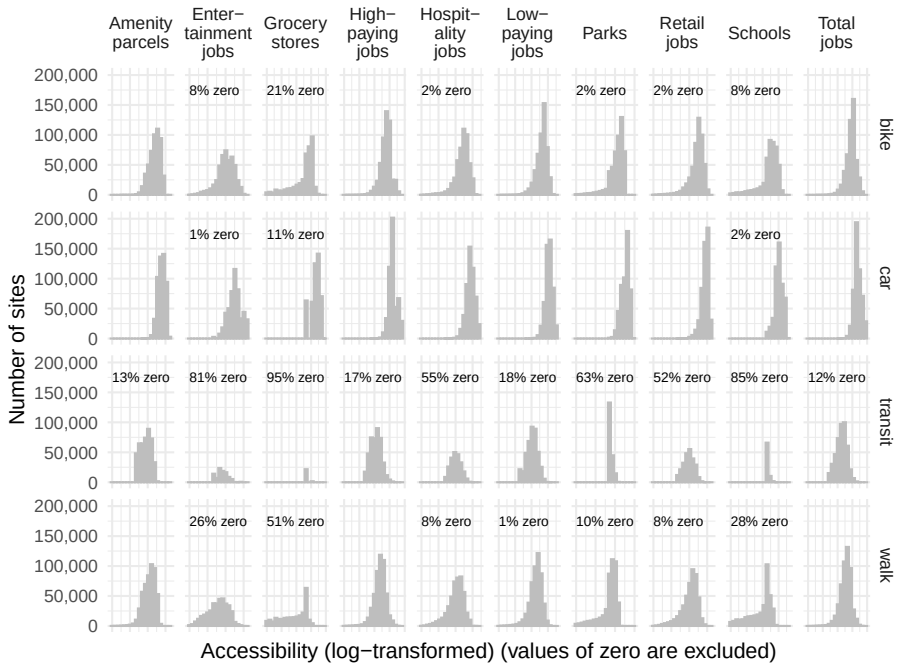

Figure 4.3: Distributions of accessibility variables

We included a disamenity proximity index in our analysis that we calculated as the logarithm of the average distance from each site to the ten closest disamenity sites. The distribution of this index is shown in 4.4.

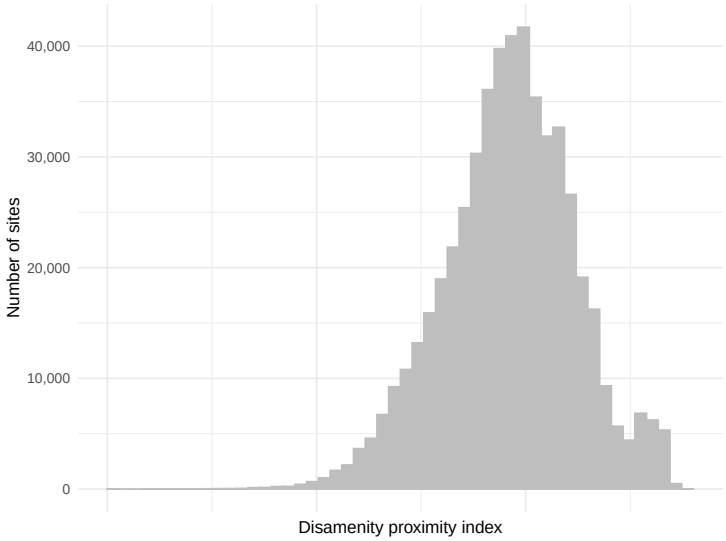

Figure 4.4: Distribution of average distance to nearest ten disamenity sites

4.1.4 Density

To represent the residential density around each site, we used the sf [Pebesma, 2018], nngeo [Dorman, 2022] and tidycensus [Walker and Herman, 2022] R packages to determine the smallest circular buffer around each site containing a population of at least two thousand people, based on the 2020 census. In denser places, a buffer with a smaller radius would encompass two thousand residents. In more sparsely-populated places, a buffer containing two thousand residents would be larger. The distribution of radii for two-thousand-person site buffers is shown in 4.5.

4.1.5 Population diversity

The two-thousand-resident buffers described above were also used as a basis to estimate the racial diversity of residents in the immediate vicinity. For each buffer, we calculated the percentage of residents that who identified in the 2020 census as non-Hispanic white, non-Hispanic Black, and Hispanic. The distributions of these variables are shown in 4.6.

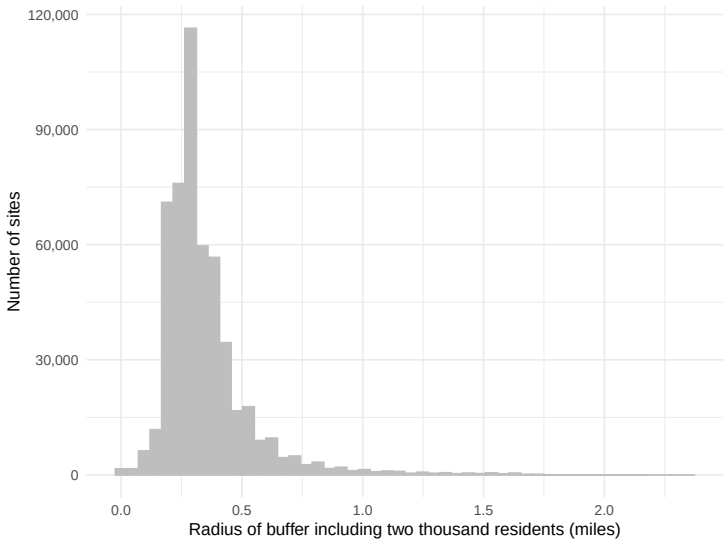

Figure 4.5: Histogram of radii of buffer containing 2000 residents

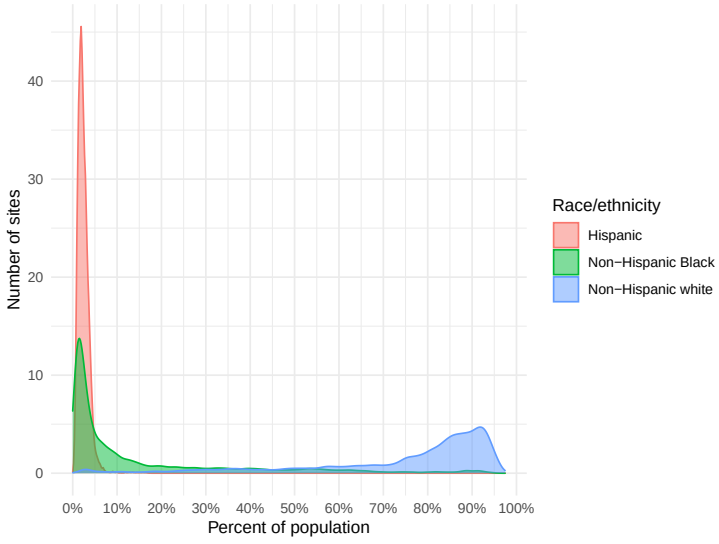

Figure 4.6: Histograms of population diversity variables

4.1.6 Land use diversity

We also calculated the total number of different land uses within each two-thousand-resident buffer and used this as a measure of land-use diversity. 4.7.

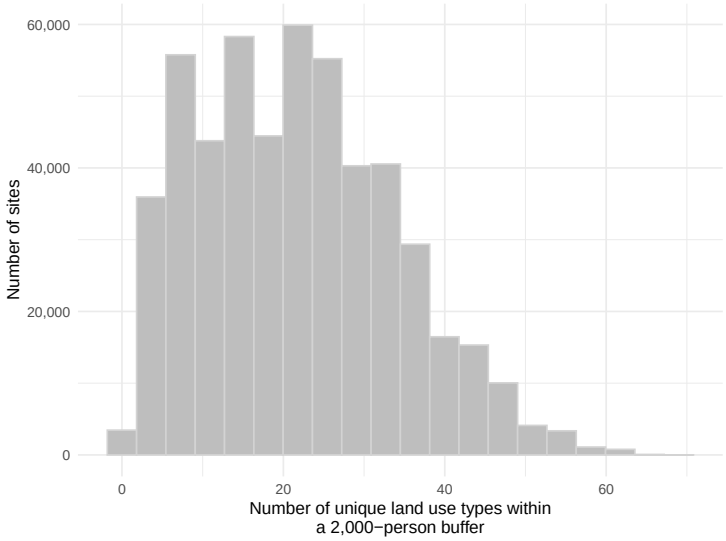

Figure 4.7: Histogram of land use diversity

4.2 Index development

The methods described above yielded a set of fifty parcel-level variables, forty of which are accessibility metrics, for each of 506,405 parcels. We used the EFAtools R package [Steiner and Grieder, 2020] to develop a set of parcel level indices from these variables using factor analysis. The Kaiser-Meyer-Olkin criterion for the dataset is 0.9, suggesting a "marvellous" case for factor analysis [Kaiser, 1974].

We determined the appropriate number of factors based on the Kaiser-Guttman criterion [Guttman, 1954] and the Hull method [Lorenzo-Seva et al., 2011] and computed factor loadings using an oblimin rotation.

Chapter 5

Results

```
library(here)
library(tidyverse)
```

5.1 Factor analysis

Both the Hull method (5.1) and the Kaiser-Guttman criterion (5.2) suggested a five-factor solution.

The loadings resulting from the factor analysis are illustrated in 5.3. We assigned names to each factor based on a visual inspection of the results. The *drivable* factor had the highest loadings for variables representing access by car to most destination types. The *walkable* factor has high loadings for variables representing access by walking and transit. The *diverse* index is characterized by diversity of people (high percentages of black residents and low percentages of white residents), diversity of land use (a greater number of distinct land uses in the immediate vicinity and a shorter average distance to disamenities), and lower assessed property values. The *dense* factor is characterized by lower values for the radius of the smallest buffer containing two thousand residents (i.e. higher population densities) and higher access to retail and grocery locations by non-motorized modes. The *amenities* factor is characterized by non-motorized and transit access to retail and grocery locations. 5.3 illustrates the loadings of each individual variable onto each of the five factors.

5.4, 5.5, 5.6, 5.7, and 5.8 show the spatial variation in the drivability, walkability, density, diversity, and amenity-richness indices, respectively.

5.9 illustrates the distribution of each factor and the relationships among them.

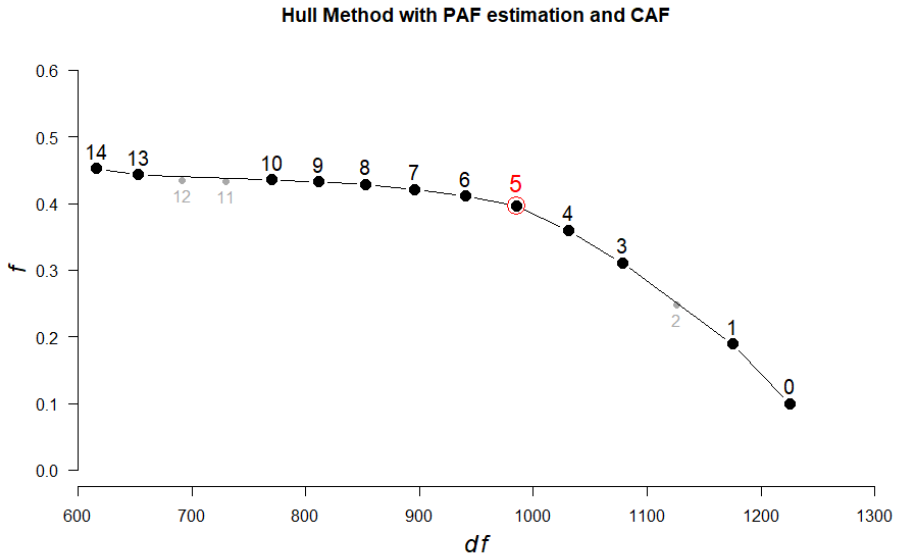

Figure 5.1: Results of Kaiser-Guttman criterion for determining the number of factors

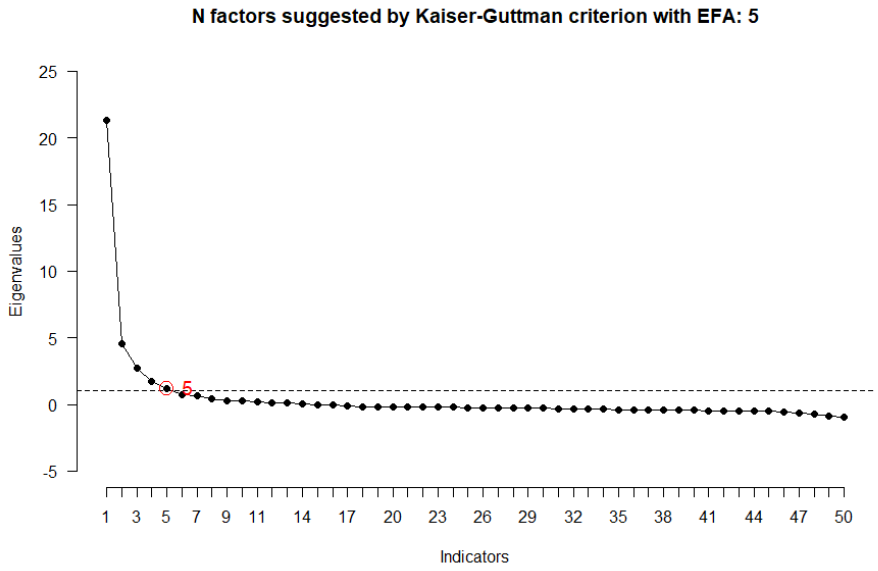

Figure 5.2: Results of Hull method for determining the number of factors

Figure 5.3: Factor loadings

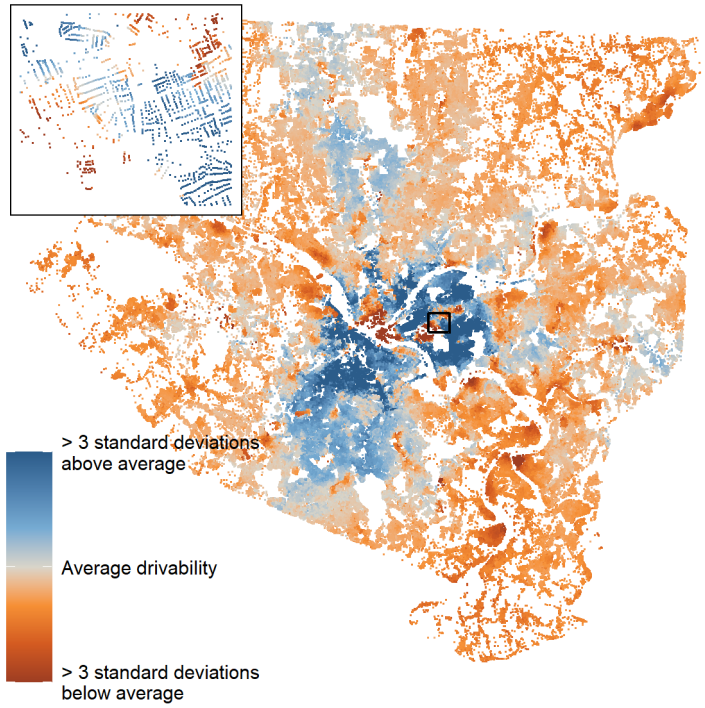

Figure 5.4: Spatial variation in drivability index

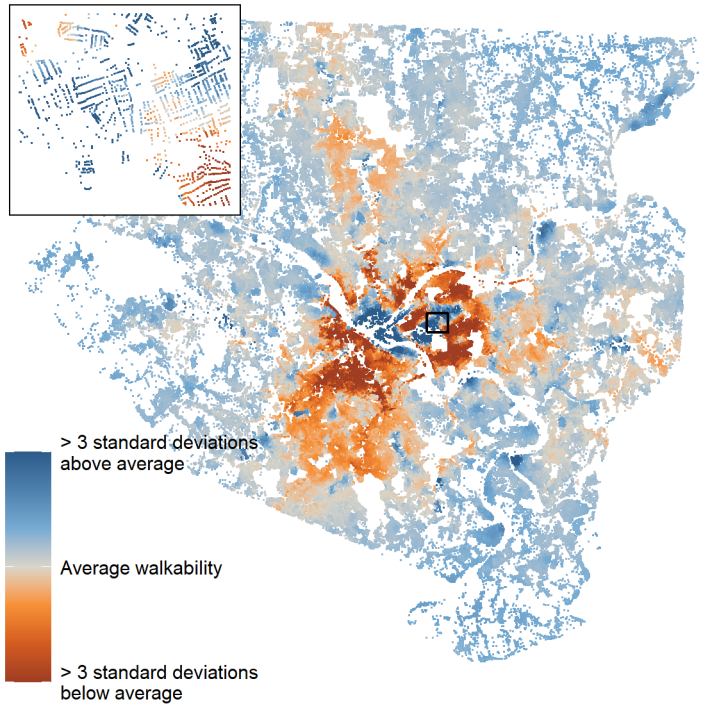

Figure 5.5: Spatial variation in walkability index

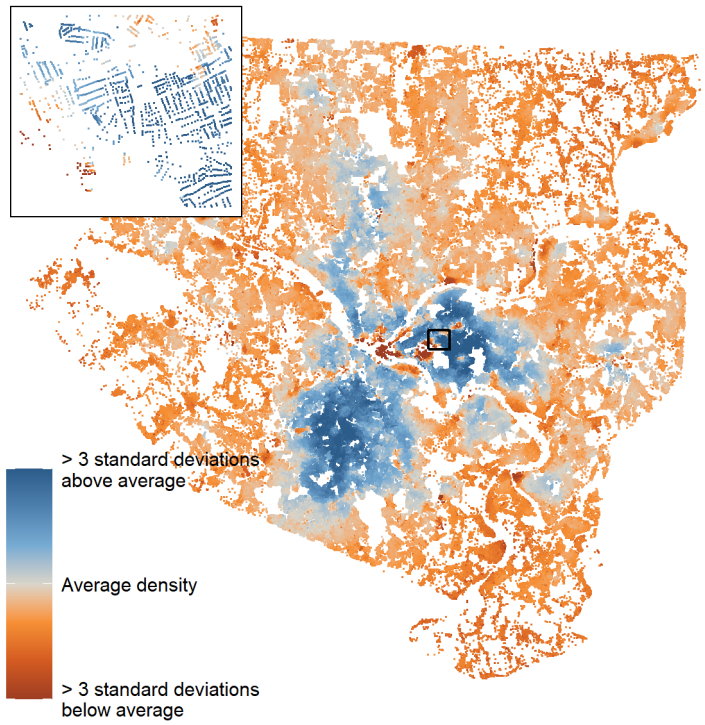

Figure 5.6: Spatial variation in density index

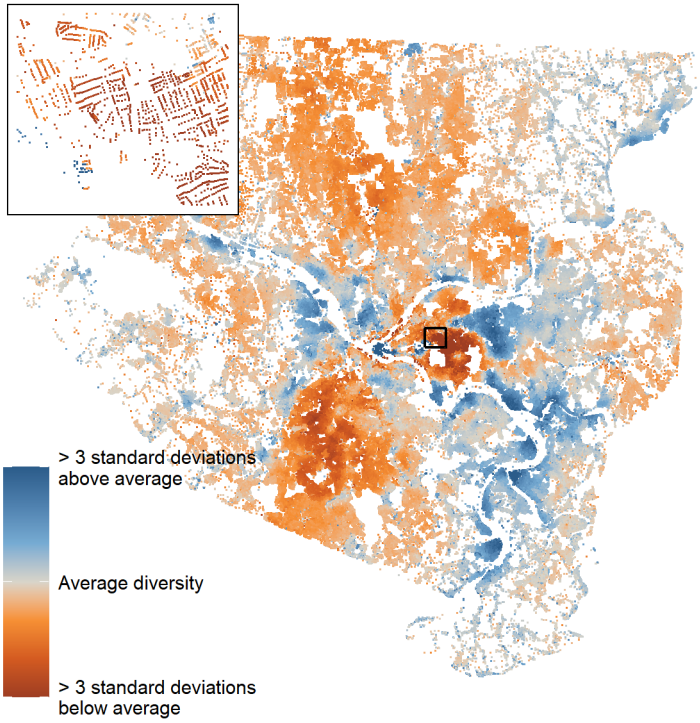

Figure 5.7: Spatial variation in diversity index

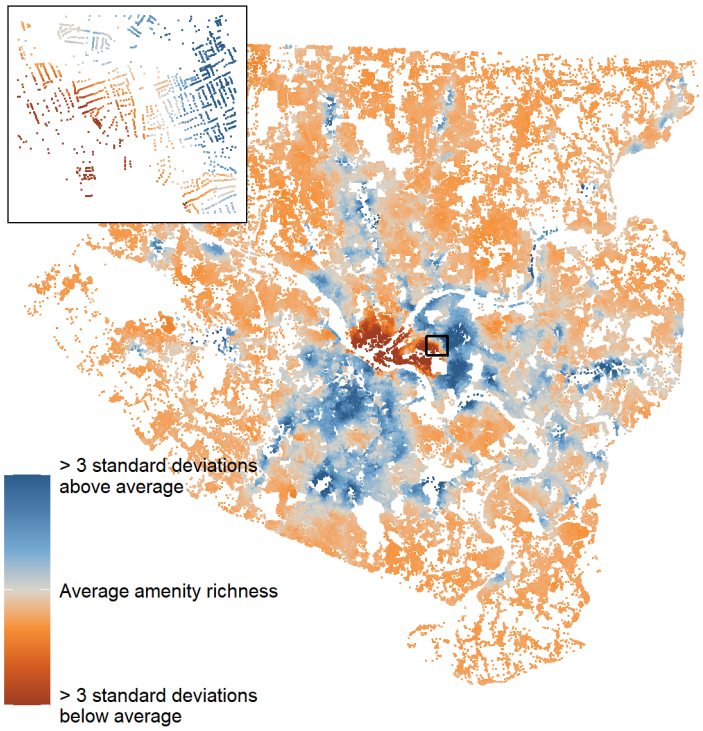

Figure 5.8: Spatial variation in amenity-richness index

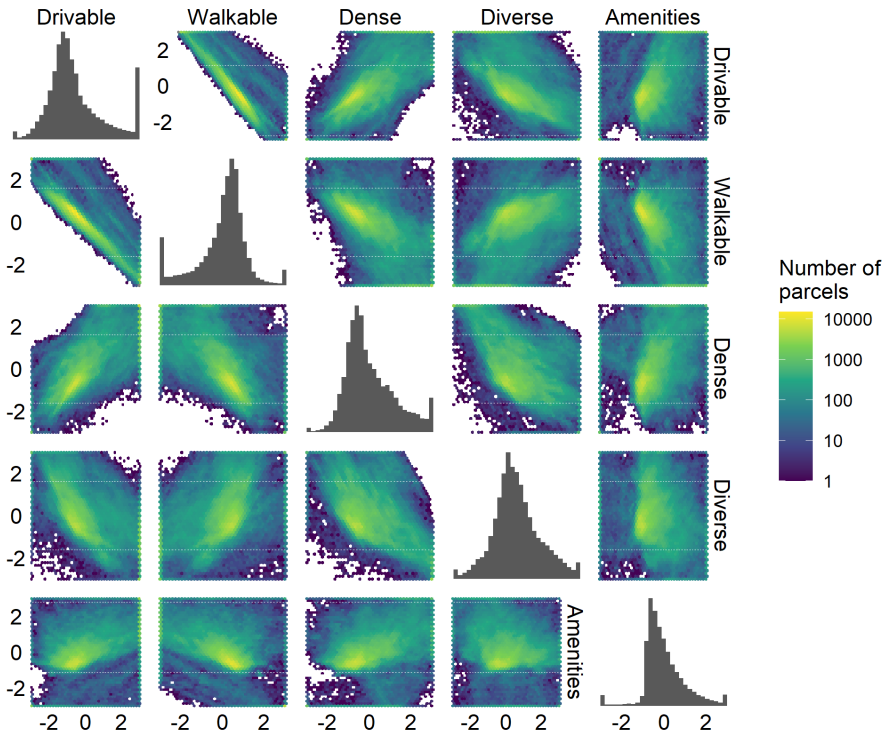

Figure 5.9: Spatial variation in amenity-richness index

Chapter 6

Blocks

6.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (6.1).

6.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 6.1.

Theorem 6.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

6.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 7

Sharing your book

7.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

7.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

7.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book- all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/gitbook.html

Or use:

```
?bookdown::gitbook
```

Bibliography

Allegheny County Office of Information Technology. Allegheny County WIC Vendor Locations, April 2018. URL https://openac-alcogis.opendata.arcgis.com/datasets/ab9ec54e46d8403db31cff6bdc890aff_0/explore?location=40.458725%2C-79.972398%2C10.20. type: dataset.

Allegheny County Office of Information Technology. Allegheny County Public Schools / Local Education Agency (LEAs) Locations, December 2020. URL https://openac-alcogis.opendata.arcgis.com/datasets/AlCoGIS::allegheny-county-public-schools-local-education-agency-leas-locations/about. type: dataset.

Allegheny County Office of Property Assessments. Allegheny County Property Assessments, May 2022. URL https://data.wprdc.org/dataset/2b3df818-601e-4f06-b150-643557229491. type: dataset.

Michael Dorman. *nngeo: k-Nearest Neighbor Join for Spatial Data*, 2022. URL https://CRAN.R-project.org/package=nngeo. R package version 0.4.5.

Louis Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161, 1954.

Henry F Kaiser. An index of factorial simplicity. *psychometrika*, 39(1):31–36, 1974.

Urbano Lorenzo-Seva, Marieke E Timmerman, and Henk AL Kiers. The hull method for selecting the number of common factors. *Multivariate behavioral research*, 46(2):340–364, 2011.

Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009.

Pennsylvania Department of Conservation and Natural Resources. Pennsylvania Local Parks Access Points, November 2015. URL https://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=308.

Rafael H. M. Pereira, Marcus Saraiva, Daniel Herszenhut, Carlos Kaue Vieira Braga, and Matthew Wigginton Conway. r5r: Rapid Realistic Routing on Multimodal Transport Networks with R^5 in R. *Findings*, page 21262, March 2021. doi: 10.32866/001c.21262. URL https://findingspress.org/article/21262-r5r-rapid-realistic-routing-on-multimodal-transport-networks-with-r-5-in-r. Publisher: Findings Press.

Markus D. Steiner and Silvia Grieder. Efatools: An r package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, 5(53):2521, 2020. doi: 10.21105/joss.02521. URL https://doi.org/10.21105/joss.02521.

United States Census Bureau. LEHD Origin-Destination Employment Statistics (LODES), October 2021. URL https://lehd.ces.census.gov/data/#lodes. Type: dataset.

Kyle Walker and Matt Herman. *tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames*, 2022. URL https://CRAN.R-project.org/package=tidycensus. R package version 1.2.

Western Pennsylvania Regional Data Center. Geocoders, February 2021. URL https://data.wprdc.org/dataset/6bb2a968-761d-48cf-ac5b-c1fc80b4fe6a.