TABLE 3.2
Expiring LIHTC Properties

	Expired 2020	Expires between 2020 and 2030	Expires after 2030	Total
HUD financing/insurance	10 (11%)	17 (18%)	66 (71%)	93 (100%)
Project-based rental assistance	147 (31%)	158 (34%)	164 (35%)	469 (100%)
LIHTC	0 (0%)	415 (28%)	1,083 (72%)	1,498 (100%)
Multiple programs	0 (0%)	3 (1%)	252 (99%)	255 (100%)

the need for targeted efforts to address this issue in these areas. Table 3.2 (data source, HUD 2022) outlines the expiration of the LIHTC units compared to other assistance programs.

3.3.6 Price Restricted Housing

In large cities like New York City, several programs actively support household stability through the provision of price-restricted housing. For example, the New York City Department of Housing Preservation and Development's Housing Development Corporation offers affordable housing lotteries in new housing developments, about 5% are set aside for residents with mobility impairments and 2% for residents with visual and hearing impairments. This process is managed through an online process; the *Housing Connect web portal* allows prospective residents to create a profile and apply for a housing lottery. If selected and eligibility is confirmed, the resident can sign a rental lease or complete a purchase agreement. There is an income cap for these lotteries. Privately owned buildings have both rental and ownership opportunities. Ownership opportunities are typically in the form of a cooperative. Eligibility includes 12 months of positive rental history and meeting income requirements (for a family of 4, eligible incomes range from $0 to $220,110). The rent paid at the winning lottery buildings is determined to be affordable if it is below 33% of the individual's annual income. The program is designed for a wide range of household sizes and income levels. The income eligibility is from 0% to 30% of the federal area median income (AMI) to 165% AMI (NYC Housing Preservation and Development HPD, 2023).

The Section 32 Homeownership Program is a federal policy that allows first time homebuyers who are at or below 80% of the AMI to receive a 20% discount of home's appraised value, along with guidance to navigate the home-buying process and may include grants to cover down payments and closing costs, a one-year home warranty and lower monthly payments.

HUD's Scattered Sites Housing Programs have been in place for over five decades, serving to disperse and deconcentrate public housing in cities with dense public housing clusters. These programs create low density housing (generally under fifteen units) in middle-income neighborhoods. Scattered Site programs can be managed by city agencies and nonprofit organizations. Many cities have successfully used scattered sites programs to create permanent supportive housing. Unfortunately, there is more demand for housing than supply creating long wait times.

The Mitchell-Lama program is a unique program serving both New York State and New York City. It is intended to create affordable housing for middle class households, and is named after its sponsors, two elected officials, State Senator Mitchell and Assembly person Lama who established the program in 1955. The program has been very successful in creating stable affordable housing through cooperatives and rentals. The original program is no longer active, but it is estimated that over 100,000 apartment units were created as a result of the program. Developers were able to delist their apartments from participation in the program after a 20-year period which impacts the availability of affordable housing, see Figure 3.19 (NYCHPD, n.d.).

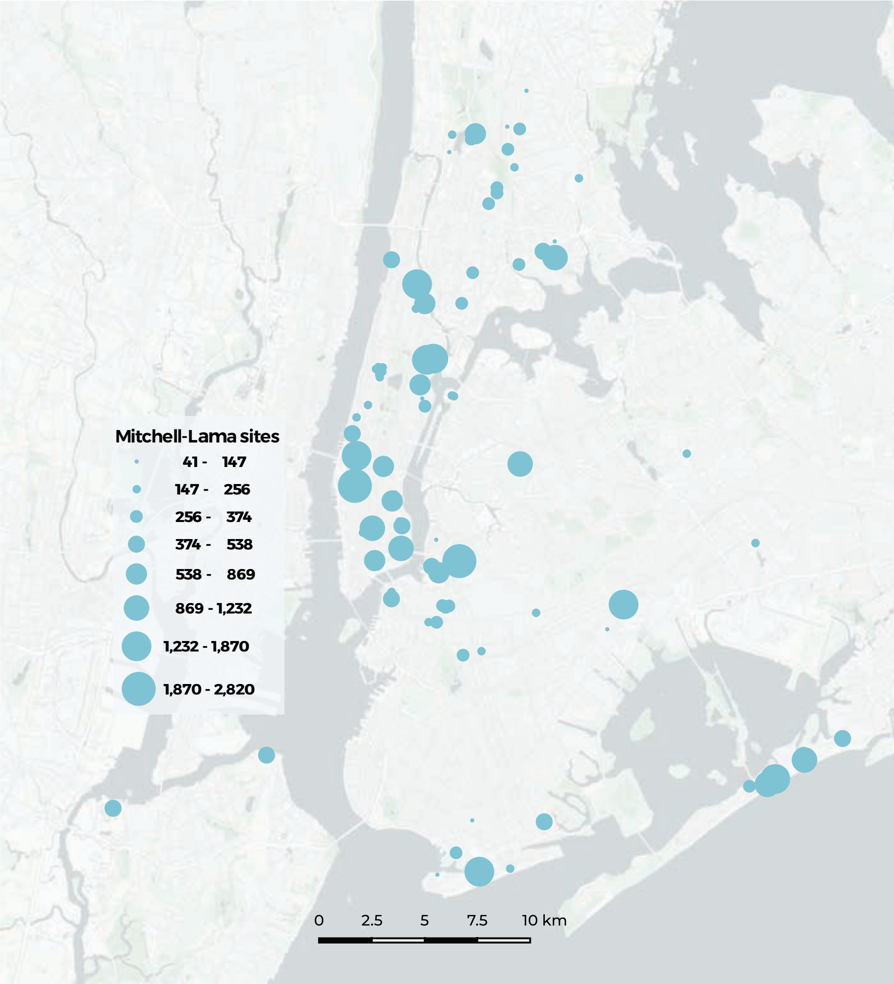


FIGURE 3.19 Mapping Mitchell-Lama development sites in New York City

3.4 EMERGING TRENDS IN ALTERNATIVE HOME OWNERSHIP – SHARED EQUITY HOMEOWNERSHIP

Condominiums, cooperatives, and community land trusts are forms of shared equity homeownership. These models provide mechanisms that make it easier to avoid speculative practices in real estate which can drive up housing prices and create instability. Condominiums began in the 1960s in Manhattan high-rise buildings. By sharing the ownership and responsibilities, condominiums, cooperatives, and land trusts create a community for the betterment of the entire building or complex. The boards or associations that govern these shared equity communities are responsible for looking after the interests of the whole, ensuring maintenance, management, and financial stability of the properties.

Today, new models for shared equity homeownership are emerging in the form or resale-restricted, owner-occupied housing, community land trusts (CLTs), limited equity cooperatives (LECs), and price-restrictive houses and condominiums with 30 plus years affordability covenants (Davis, 2018). These homeownership models are an alternative to single-family homes, offering diverse housing options and an opportunity for individuals and families to become homeowners while benefiting from shared amenities, reduced maintenance costs, and a cooperative living environment.

3.4.1 Condominiums and Cooperatives

Condominiums and cooperatives have many similarities but are different in their ownership structure and the rights and responsibilities of the residents. A unit owner of a condominium has direct ownership of their unit and holds a deed to the unit directly. Common areas are owned collectively by the unit owners through an association or cooperation. Decisions in a condominium unit are made by individual owners but must operate within the rules set by the condo associations bylaws. Alternatively, members of cooperatives, or coops, do not own their individual unit. They own shares or memberships in the coop corporation, which owns the entire building. Each resident holds a lease, which allows them to live in a specific unit. Coop residents have voting rights and can participate in its governance. Major decisions are made collectively by a co-op board of directors or general assemblies, where residents have the opportunity to voice their opinion and vote on important matters.

The structure of a condominium and cooperative may evoke high-rise buildings in New York City. Figure 3.20 (MapPLUTO, n.d.) shows the square footage of condominiums in New York City. Manhattan has a high density of large condominium buildings with very large square footage, while the outer boroughs have both these large buildings as well as many smaller buildings. While high-rise buildings are prevalent in cities like NYC, cooperative (coop) and condominium (condo) housing typologies can also be found throughout the United States in the form of semi-detached townhouse buildings within expansive complexes, often gated communities. These communities provide various amenities like community pools, fitness centers, tennis courts, or golf courses. Given that 88% of all housing structures in the United States consist of attached or detached single-family homes, opting for shared equity homeownership models such as coops and condos presents an alternative route to

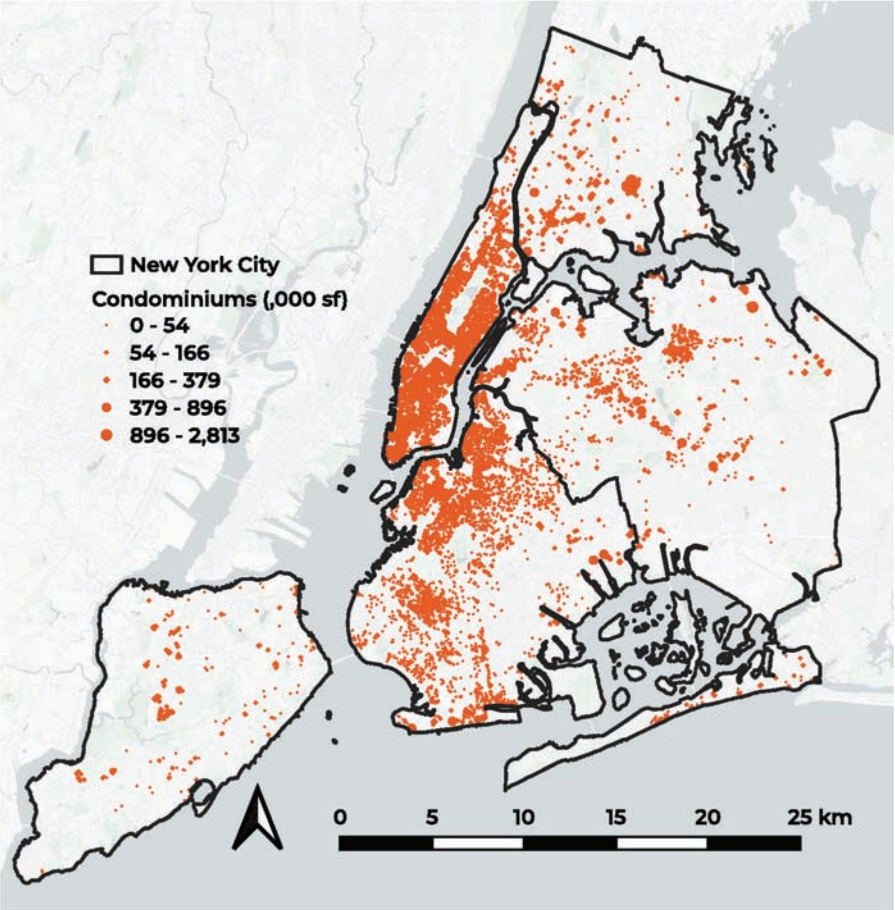


FIGURE 3.20 Mapping the concentration of condominiums in New York City

homeownership that may be more feasible for those who would otherwise find it unaffordable, see Figure 3.21 (data source, US Census 2023b).

3.4.2 Community Land Trusts

Community land trusts (CLT) are private entities that purchase property, usually in neighborhoods that have blight, in order to be able to lease land at set prices for the future. It is a “social invention designed to address social problems” (Meehan, 2014). CLT’s ownership can be made up of community residents, non-residents, and representatives with a public interest. The idea of the CLT shifts the relation of land in the hands of a private owner to that of a community. Like cooperatives, the land is owned by the CLT and leased out to individuals. The CLT idea, however, is different from cooperatives because the trust can be made up of members that are not lease holders, but rather support the social and economic goals of the CLT.

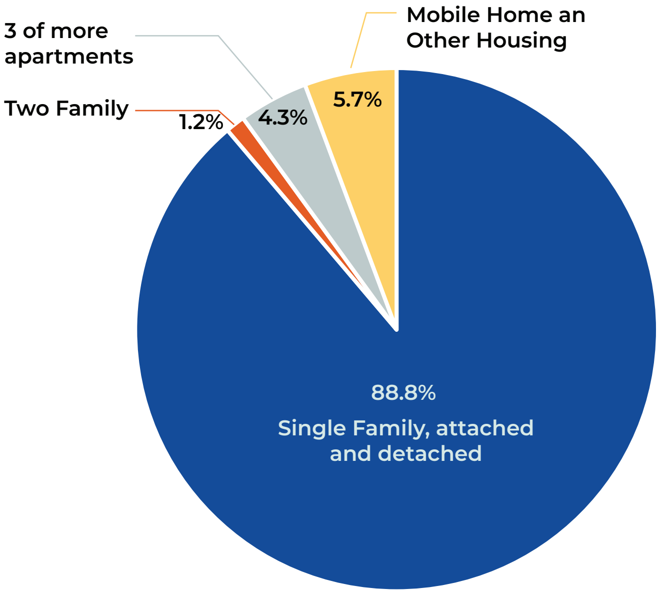


FIGURE 3.21 Percent distribution of homeowners by structure type and number of units

As of 2023, in the United States there are about 225 CLTs. These land trusts operate at different scales but benefit the members in similar ways by creating and preserving affordable housing, stabilizing communities by preventing displacement, building wealth for low-income families, and promoting ownership and control of land. In a CLT the ownership of land does not have to be contiguous and more often the locations of the land trust are scattered throughout a community, as seen in Figure 3.22 from the Oakland CLT (OakCLT Properties, 2003). Oakland CLT has lots throughout the city of Oakland and the properties in the trust range from single-family homes, transitional housing, and commercial properties.

Community land trusts (CLTs) are a growing movement in the United States, and they are playing an important role in addressing the affordable housing crisis. CLTs can help to create and preserve affordable housing, stabilize communities by preventing displacement, build wealth for low-income families, and promote community ownership and control of land. CLTs are a promising solution to the affordable housing crisis, and they are likely to play an even greater role in the years to come.

3.5 USING GIS FOR STORYTELLING AND COMMUNICATION

We have emphasized the importance of spatial relationships and the need to consider housing within its geographical context. We have heavily annotated our narrative with static maps and images to communicate specific data and evidence but also to help tell a story. Static maps allow us to compare information about spatial extents (such as county boundaries or state lines) alongside the variables under consideration – for example, average home prices at a national level mask the high variability that is

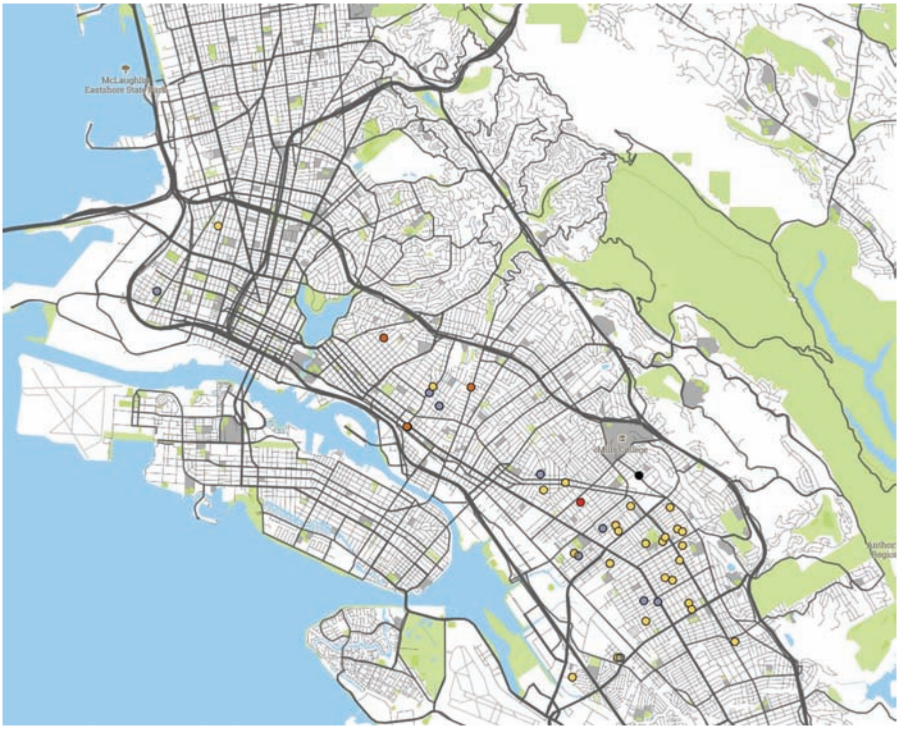


FIGURE 3.22 The disjointed spatial layout of community land trust properties in Oakland, CA

immediately apparent when visualized at the state level. Yet, static maps alone cannot fully capture the complexity of the housing challenges we encounter. Understanding relationships between different variables anchored by the same spatial extent reveals the power of using GIS analyses. However, moving from static to dynamic representations as well as the inclusion of interactive elements are powerful ways to connect with hyper-diverse audiences. The GIS company Esri has developed two products to support impactful storytelling, but the storytelling concept can be applied even without the use of the Esri tools.

StoryMaps weave together different lines of reasoning, akin to how a statistician may likely apply multivariate analyses, but use more compelling visual narratives. For example, Steven Aviles (2022) storymap (Aviles, 2022) introduces the boom and bust phases of building construction, then visualizes housing affordability at multiple geographical scales, and concludes with a discussion of policy options. Storymaps combine emotive photography with graphics and maps that are held together by a story text of, in this case, some 2,000 words. In the Aviles example above, the section on construction history consists of three maps, one each for the westward expansion (similar to our Figure 2.2), urban sprawl (compare Figure 1.13), and first and second ring of suburbs (see our Figure 2.25). The section on affordability is supported by a table, five charts, and no less than 36 maps of eight metropolitan areas. The story ends with a discussion of zoning, ADUs and LIHTC illustrated with a graphic explaining policies aimed at densification like California's 2020 bill on subdivisions, known as *SB 1120*. Since storymaps are dynamic web pages, it is easy to enrich them with external links and embedded content.

The author of this very comprehensive storymap had the advantage of being able to work with detailed nationwide datasets compiled by their employer Esri. This allowed him to create uniform maps for the eight metro areas without having to go through the data assembly strategies that we are going to discuss in Chapter 4. A more typical storymap would tell the same story for just one study area such as *Madison's (WI) storymap explaining densification*, and it would be fairly straightforward to create if the housing researcher accesses the data that is usually held in-house in a local or regional authority. We will revisit the notion of storymaps in Chapter 5, where we expand on their ability to communicate complex analyses to local audiences. While well hidden under the shiny presentation, each storymap relies on the same data that we will use in Chapter 5 to introduce the reader to GIS analyses. The domain of housing-related data is rich and not quite self-explanatory, and for this reason, we have devoted a whole chapter on data for housing research.

NOTES

1. Contrary to the popular image of a stand-alone house in the middle of a yard, single-family homes also include condominiums and townhouses.
2. Some states have now (2023) started to propose guidelines for fire safety in tiny homes.

REFERENCES

ADU Handbook, 2022. *HCD Accessory Dwelling Handbook*. Sacramento, CA: California Department of Housing and Urban Development. https://www.hcd.ca.gov/sites/default/files/2022-07/ADUHandbookUpdate.pdf, last accessed 5 May 2023.

Alexander, L, 2022. "Tiny Homes: A Big Solution to American Housing Insecurity". *Harvard Law & Policy Review*, 15(2): 471–509. https://scholarship.law.tamu.edu/facscholar/1551, last accessed 21 May 2023.

Aviles, S, 2022. *How the Age of Housing Impacts Affordability*. https://storymaps.arcgis.com/stories/ae7f226a5ffd4466acbe0c7a14deab0e, last accessed 27 May 2023.

Calavita, N, and Grimes, K, 1998. "Inclusionary Housing in California: The Experience of Two Decades". *Journal of the American Planning Association*, 64(2): 150–169. https://doi.org/10.1080/01944369808975973

Calthorpe, P, and Fulton, W, 2001. *The Regional City: Planning for the End of Sprawl*. Washington, DC: Island Press.

Community Loan Fund, 2023. *New Hampshire Community Loan Fund*. https://communityloanfund.org/, last accessed 5 May 2023.

Davis, J, 2018. "More than Money: What Is Shared in Shared Equity Homeownership?" *Shelterforce*, 156(2): 30–33.

Federal Transit Administration (FTA), 2023. *Transit Oriented Development*. https://www.transit.dot.gov/TOD, last accessed 31 May 2023.

Gans, H, 1967, 2017. *The Levittowners: Ways of Life and Politics in a New Suburban Community*. New York: Columbia University Press.

Google Maps, 2023, last accessed 21 October 2023.

HUD, 2021. *Understanding SRO*. Washington, DC: US Department of Housing and Urban Development. https://files.hudexchange.info/resources/documents/Understanding-SRO.pdf, last accessed 15 May 2023.

HUD, 2022. *The Low-Income Housing Tax Credit (LIHTC)*. Washington, DC: US Department of Housing and Urban Development. https://www.huduser.gov/portal/datasets/lihtc.html, last accessed 15 May 2023.

HUD, 2023a. *Rental Assistance Demonstration: Conversion Guide for Public Housing Agencies*. Washington, DC: US Department of Housing and Urban Development. https://www.hud.gov/sites/documents/RADCONVERGUIDEPHA.PDF, last accessed 15 May 2023.

HUD, 2023b. *The Office of Manufactures Housing Programs*. Washington, DC: US Department of Housing and Urban Development. https://www.hud.gov/OMHP, last accessed 31 May 2023.

HUD User, 2023. *Low-Income Housing Tax Credit (LIHTC): Property Level Data*. https://www.huduser.gov/portal/datasets/lihtc/property.html, last accessed 5 May 2023.

LEED, 2023. *LEED v4: Reference Guide for Neighborhood Development*. Washington, DC: U.S. Green Building Council. https://www.usgbc.org/guide, last accessed 28 May 2023.

Lowenkron, H, 2021. *Creating More Accessible Inclusive Buildings*. New York: Bloomberg City Lab + Equality. https://www.bloomberg.com/news/features/2021-08-18/how-universal-design-creates-inclusive-infrastructure, last accessed 21 May 2023.

Manufactured Housing Institute (MHI), 2023. *2023 Manufactured Housing Facts: Industry Overview*. https://www.manufacturedhousing.org/wp-content/uploads/2023/06/Industry-Overview.pdf, last accessed 11 May 2023.

MapPLUTO, n.d. *New York City Department of Urban Planning and Development parcel-level database*. Online resource available at https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.

McCammant, K, and Durrett, C, 2011. *Creating Cohousing: Building Sustainable Communities*, 3rd edition. Gabriola Island, BC: New Society Publishers.

Meehan, J, 2014. “Reinventing Real Estate: The Community Land Trust As a Social Invention in Affordable Housing.” *Journal of Applied Social Science*, 8(2): 113–133.

Mukhopadhhyay, J, Ore, J, and Mende, K, 2019. “Assessing Housing Retrofits in Historic Districts, in Havre, Montana”. *Energy Reports*, 5: 489–500, doi:10.1016/j.egyr.2019.03.008

Museum of the City of New York: Byron Collection, in World History Commons, https://worldhistorycommons.org/museum-city-new-york-byron-collection, last accessed 12 July 2023.

Nadel, S, Prindle, B, and Brooks, S, 2005. *The Energy Policy Act of 2005: Energy Efficiency Provisions and Implications for Future Policy Efforts*. Washington, DC: American Council for an Energy-Efficient Economy.

NYCDCP, n.d. *Inclusionary zoning developments in NYC. New York City Department of City Planning Inclusionary Housing Designated Areas*. Online resource available at https://data.cityofnewyork.us/City-Government/Inclusionary-Housing-Designated-Areas/w83z-2kf9.

NYCHPD, n.d. List of Mitchell-Lama addresses. New York City Department of Housing Preservation and Development. Online resource available at https://www.nyc.gov/assets/hpd/downloads/pdfs/services/MLLIST.pdf, last accessed 05/17/2023.

NYC Housing Connect, 2023. *Housing Connect Web Portal*. https://housingconnect.nyc.gov/PublicWeb/, last accessed 15 May 2023.

NYC Housing Preservation and Development, 2023. *Affordable Housing*. https://www.nyc.gov/site/hpd/services-and-information/find-affordable-housing.page, last accesses 15 May 2023.

OakCLT Properties. *OakCLT Properties*. 2023. https://oakclt.org/about/oakclt-properties/, 13 July 2023.

Parolek, D, 2020. *Missing Middle Housing: Thinking Big and Building Small to Respond to Today's Housing Crisis*. Washington, DC: New YorkIsland Press.

RAD, 2023. *Properties Participating in RAD Program*. RAD Resource Desk. https://www.radresource.net/pha_data.cfm, last accessed 5 May 2023.

Sanguinetti, A, 2015. "Diversifying Cohousing: The Retrofit Model". *Journal of Architectural and Planning Research*, 32(1): 68–90.

Scally, C, Gold, A, and Dubois, N, 2019. *The Low-Income Housing Tax Credit: How it Works and Who It Serves*. Washington DC: The Urban Institute.

Smith, R, and Bereitschaft, B, 2016. "Sustainable Urban Development? Exploring the Locational Attributes of LEED-ND Projects in the United States through a GIS Analysis of Light Intensity and Land Use". *Sustainability*, 8: 547. doi:10.3390/su8060547

SRO Housing Corporation. *SRO Housing Corporation*. 2023. https://www.srohousing.org/property-management.html, 13 July 2023

Talen, E, 2005. *New Urbanism and American Planning. The Conflict of Cultures*. Philadelphia, PA: Routledge.

US Census Bureau, 2017. *American Community Survey - Data Tables and Tools - Subject Tables*. Online Source https://www.census.gov/acs/www/data/data-tables-and-tools/subject-tables/, 28 July 2023.

U.S. Census Bureau, 2021. *2020 Census Group Quarters*. https://www.census.gov/newsroom/blogs/random-samplings/2021/03/2020-census-group-quarters.html#:~:text=Group%20quarters%20are%20defined%20as,not%20related%20to%20one%20another.

U.S. Census Bureau, 2023a. *Glossary*. https://www.census.gov/glossary/, last accessed 31 May 2023.

U.S. Census Bureau, 2023b. *Housing Vacancies and Homeownership*. https://www.census.gov/housing/hvs/index.html, last accessed 31 May 2023.

U.S. Census Bureau, 2023c. *Latest Data Tables of New Manufactured Homes*. https://www.census.gov/data/tables/time-series/econ/mhs/latest-data.html, last accessed 26 May 2023.

U.S. Census Bureau, 2023d. *U.S. Census Bureau Construction Spending - Characteristics*. n.d. Online Source https://www.census.gov/construction/chars/current.html, 28 July 2023.

U.S. Census Bureau and U.S. Department of Housing and Urban Development (HUD), Houses Sold by Type of Financing, Cash Purchase [HSTFC], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/HSTFC, 20 October 2023.

USHUD, n.d. US Department of Housing and Urban Development LIHTC property level data in MS Access and CSV format. Online resource available at https://www.huduser.gov/portal/datasets/lihtc/property.html.

Zhao, D, McCoy, A, Agee, P, Mo, Y, Reichard, G, and Paige, F, 2020. "Time Effects of Green Buildings on Energy Use for Low-Income Households: A Longitudinal Study in the United States". *Energy Policy*, 140: 111827. doi:10.1016/j.enpol.2020.111827.

4 Data for Housing Research

4.1 HOUSING DATA SOURCES

Most housing data is collected by organizations that have a financial stake in housing and need the data for the purposes of financial accountability. As such, housing data is generated (although not necessarily published) by everyone who has a financial stake in the housing market: lenders, insurances, builders, private, cooperative, governmental, or non-governmental entities. In addition, there are some data collections by foundations, think tanks, and academic institutions, although they are more often than not ad hoc; i.e., they tend to be compilations for a particular study rather than long-term repositories. The quintessential counterpart to these is the US Census Bureau, which has been collecting housing-related data for almost a century. Table 4.1 in the Appendix provides an overview of the range of suitable data sources.

4.1.1 US Census

At the moment, much of housing policy analysis conducted by planners focuses on analysis at the state level, comparing the impacts of government policies in different states, for instance, or at the level of level of counties. There are a little over 3,000 counties in the United States, and over 84,000 census tracts! Counties can be large or small, and often county-level analysis cannot provide the fine-grained spatial differentiation of phenomena that is necessary to understand policy or programmatic impacts.

In a very narrow sense, the responsibility of the US Census Bureau is to enumerate the population of the United States every 10 years for the purpose of apportionment of seats in the House of Representatives (ref). Given the size of the task, the Census Bureau harbors a large number of experts in the fields of demography and statistics. This in turn led to the request of many other government agencies to use these resources for the collection of a wide range of other data. As the core counting unit of the census is a household at a given address, housing is the next logical realm of data to be collected.

Across a myriad of censuses and surveys, the US Census Bureau collects literally thousands of variables, many of which are useful for housing policy research. What makes Census data quintessential GIS data, however, is the fact that each data point has a spatial reference, i.e., it refers to an area unit that is both unique and well specified. The Census Bureau is by law required to preserve confidentiality about the data collected, which means that for 72 years after each collection, data is published in aggregate form only. There are multiple ways that individual-level data can be aggregated and the result is an interesting relationship between spatial, temporal,

DOI: 10.1201/9780429243066-4

TABLE 4.1
Data Source List/Summary

US Census Bureau	Information on homeownership rates, housing vacancies, and housing market characteristics
Zillow	Data on home values, rental prices, and other housing-related information
Redfin	Data on home sales, prices, and market trends. They also provide an API for accessing their data
Realtor.com	Real estate listings, property information, and housing market data
Federal Housing Finance Agency (FHFA)	Data on home prices, mortgage rates, and mortgage market conditions. They also maintain the House Price Index (HPI), which tracks changes in home prices over time
National Associations of Realtors (NAR)	Regular reports on existing home sales, home prices, and housing market trends in the United States
Bureau of Economic Analysis (BEA)	Data on housing investment, construction spending, and other economic indicators related to the housing market
Department of Housing and Urban Development (HUD)	Information on affordable housing programs, housing market conditions, and demographic data
CoreLogic	Real estate market information, including property values, mortgage data, and housing market trends
Local Multiple Listing Services (MLS)	Regional or local databases used by real estate agents to list and share property information
Home Mortgage Disclosure Act (HMDA)	Information on mortgage lending, including loan types, interest rates, and borrower demographic
National Association of Home Builders (NAHB)	Housing market data, including home construction statistics, building permits, and industry trends
S&P Case-Shiller Home Price Indices	Data on home prices in major metropolitan areas across the United States
Federal Reserve Economic Data (FRED)	Economic and housing-related data, including housing starts, building permits, and mortgage rates
Urban Institute	Datasets related to affordable housing, housing market dynamics, and housing finance
Mortgage Bankers Association (MBA)	Data on mortgage applications, refinancing activity, and mortgage market trends
Federal Housing Administration (FHA)	Data on government-insured mortgage loans, including loan volumes, delinquency rates, and borrower demographics
Local and regional government websites	Information on property taxes, housing permits, and neighborhood statistics

and attribute specificity requiring end users to make some choices of how to set up their data queries: Census data can be very detailed but would then be representative only for large areas and somewhat outdated, or can be very specific to a subset of a neighborhood but only for a few common variables and again at the price of low currency, or it can be collected every month but only at the spatial resolution of counties,

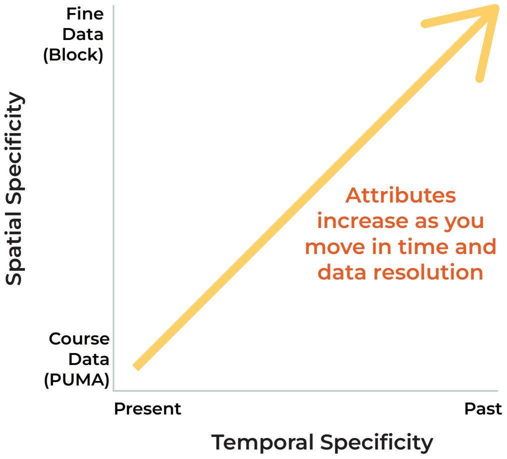


FIGURE 4.1 Choosing between spatial, temporal or attribute specificity

i.e., some 3,000 data points for the whole country. Figure 4.1 describes this conundrum of how to navigate between the three opposing characteristics.

The temporal resolution of Census data ranges from being reported monthly (e.g., employment statistics), to yearly, 3-yearly, 5-yearly, and only once in a decade. The spatial resolution is more complicated as illustrated in Figure 4.2. The lowest level of aggregation (or in Census parlance summary level) is a census block, an area unit that on average captures some 400 people and aims to be delineated by topographic features such as a street block. Only a few very common variables are released once every 10 years at this fine spatial grain. Typically, three or four census blocks are then aggregated to establish block-groups and some variables collected over a span of 5 years are published at this level. Most, though not all, variables are available at the next higher level of spatial aggregation, the census tract. And so it goes up the ladder of Figure 4.2. All the area units along the central spine of this figure fit neatly into each other, i.e., their boundaries never intersect or cross. As more and more other government agencies asked for aggregations according to their needs, the Census Bureau also publishes data in area units such as municipal, school district, or ZIP code area boundaries, in other words – special purpose boundaries that are useful for management and governance.

In addition to the decadal Census of Populations and Households that was mentioned previously, the Census Bureau conducts a continuous American Community Survey (ACS), the results of which are published in 1-, 3- and 5-year intervals (aggregates) with gradually increasing levels of spatial specificity as the data is aggregated over longer time spans. While a census aims to be a complete enumeration of all entities of its universe (here, people or households), a survey (even one as large as the ACS) is based on a sample of the statistical population and results are estimates. Therefore, all ACS data releases are accompanied by a reference to confidence ranges. Each of the variables of these products is independent; i.e., if we have small area data for a specific time span that provides information about income and rents then we *cannot* combine these variables to deduce a causal link between these variables, that is we cannot establish the number of people in one income group category that pays a particular amount of rent. The Census Bureau

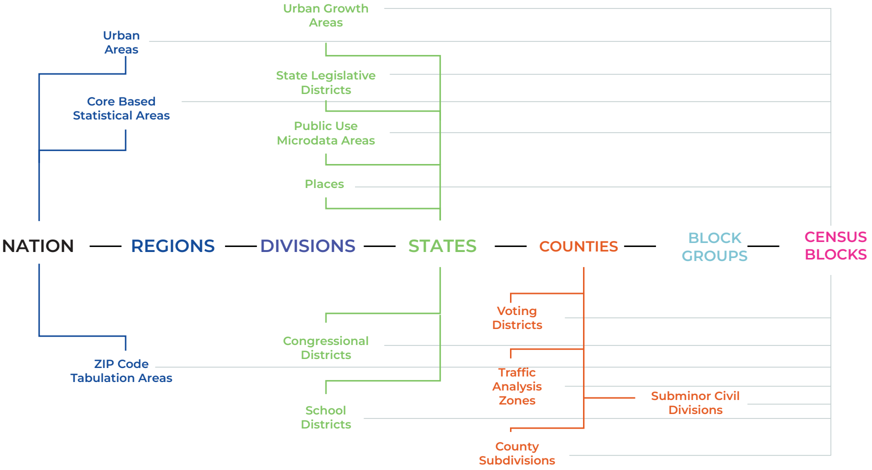


FIGURE 4.2 Hierarchy of Census area units

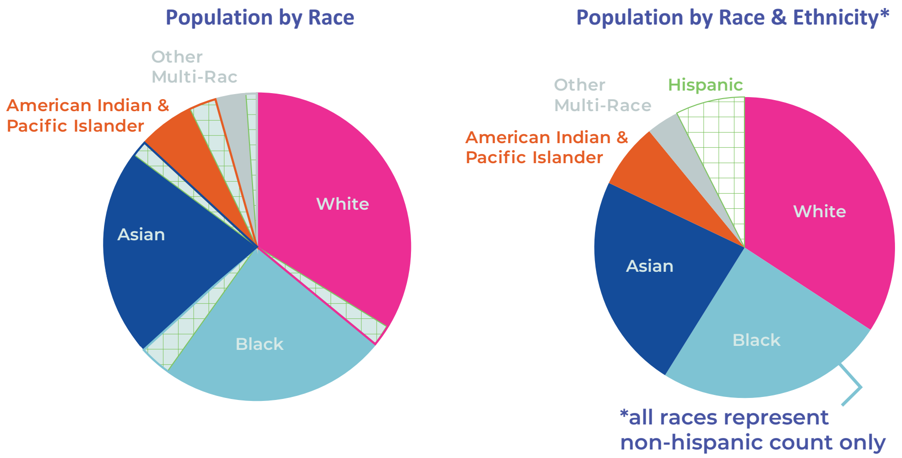
does perform such calculations based on individual-level data, but such combinations of variable values are then only released at much coarser spatial resolutions (so-called Public Use Microdata Areas (PUMA) are designed in such a way that each PUMA has no less than 100,000 but often as many as 200,000 people). In addition to these very large and comprehensive products, the Census Bureau conducts many dozens of other more specialized data collections such as the American Housing Survey, Consumer Expenditure Survey, Housing Vacancy Survey, Annual Business Survey, Annual Survey of Public Employment and Payroll, Annual Survey of State and Local Government Finances, Building Permits Survey, the Census of Governments, the Economic Census, or the Survey of Construction among many others. The sheer volume of data makes the Census website somewhat difficult to navigate. Dedicated third-party websites that transform Census datasets to make them accessible for diverse audiences include *Social Explorer*, Esri's *Living Atlas,* or the *Census Reporter* discussed in the following section.

4.1.2 Census Reporter and Social Explorer

The US Census Bureau's web site requires a good understanding of the types of data collections, the area units, and the intricacies of attributes. Many of the datasets are unwieldy, containing hundreds of columns. This is great for expert users, who typically use application programming interfaces (APIs) to access the data they need quickly. Casual or novice users tend to get intimidated. To serve these constituents and to create access and equity, non-for-profit organizations and academic institutions have created web sites that provide the user with the results of commonly run queries and reformat the output into easily digestible spreadsheets and a number of exportable GIS formats such as KML, GeoJSON, and Geopackage. It is useful to note that these efforts have been underway for many decades and the data offerings, and the data provided have co-evolved with hardware and software advances.

Ethnicity	Total Count	Hispanic Count
White	411	20
Black	312	30
Asian	270	5
American Indian & Pacific Islander	100	20
Other Multi-Race	50	10

Each Race has a % of Hispanic Ethnicity. To include Hispanic Population on the same chart you can remove the Hispanic Count from each race category. The race categories then become non-hispanic "race"


Population by Race

Population by Race & Ethnicity*

*all races represent non-hispanic count only

FIGURE 4.3 Misinterpreted identities: unraveling the consequences of misaligned race and ethnicity data reporting

The *Census Reporter* is a non-profit organization that strives to make data from the American Community Survey easier to use. The Knight Foundation funded the initial build-out of the site, which is now maintained by Northwestern University's School of Journalism and hosted by Oregon State University. In addition to precompiled profiles for over 20 topics, the site provides tutorials on the Census geographies, table organization, and technical background that help site visitors to slowly transition from the *Census Reporter* to work directly with the US Census website. For each of the topic areas, the site provides not only data but means to generate graphics and maps, which web site visitors can then download or embed in their own website.

Let us consider the one topic area of interest to us: Housing! The *Census Reporter* describes it as follows.

> The American Community Survey gathers extensive data about the housing conditions of respondents, including whether they own or rent their home, how much they spend on housing, and the physical characteristics of homes. Most of the tables count the number of housing units for a given characteristic. However, a few tables estimate the number of people living in owned or rented housing units. A housing unit is anything from a house to an apartment or even a boat if a person is currently living there
>
> *(US Census 2020 FAQ).*

Every housing unit is recorded as either occupied or vacant. Some vacancies are market related, such as houses for sale or apartments for rent. Other housing units are seasonally vacant. Occupied housing units in the ACS are split into two categories: renter-occupied and owner-occupied. This distinction is known as tenure.

The appendix contains a number of lengthy tables that illustrate the sometimes overwhelming wealth of ACS data. As we will discuss in some detail in Section 4.5, the selection of data should be based on one's conceptual model and research question. For example, we may want to look at measures of neighborhood stability. If this is the case, one of the first ACS variables to look at would be geographic mobility. Conventional wisdom has it that rented housing units see more of a turnover than owned properties – with associated assumptions about housing quality or even crime. But is this true? In New York City, for instance, there are rent-stabilized neighborhoods that result in tenants staying for many decades while gentrifying neighborhoods experience significant amounts of flipping, i.e., buyers purchase the property as a real estate investment instead of a residence. The ACS provides us with a number of variables in both tenancy categories that help us to investigate the question, and the answer is of course varying from one real estate market to another – often even within a single county or city. In addition to the geographic mobility variable, which can be reverse-interpreted as what percentage of an area unit's population has been staying in place for a certain number of years, we could look at mortgage status (Table 4.3), where a low number of mortgaged properties are either a function of an old (and stable) housing stock or of flipping (which is financially more lucrative when the property is purchased with cash, thereby avoiding the interest costs). Stable residential neighborhoods are marked by stable home values, i.e., no rapid value changes when compared to those in the vicinity. The ACS housing value variables are given in Table 4.2. Housing affordability is not well captured by mere rent or purchasing costs. Both tenancy types have associated costs such as maintenance, utilities, insurance, taxes, etc. Each of these may (but don't necessarily do) add significantly

TABLE 4.2
NHGIS GIS File Availability

	2020	2012–2019	2011	2010	2009	2000	1990	1980	1950–1970	1910–1940	1790–1900
Nation	X	X	X	X	X	X					
Region	X	X	X	X	X	X					
Division	X	X	X	X	X	X					
State	X	X	X	X	X	X	X	X	X	X	X
County	X	X	X	X	X	X	X	X	X	X	X
Census tract	X	X	X	X	*	X	X	X	X	X	
Block group	X	X	X	X	*	X	X				
Block	X			X		X	X	X			

Source: https://www.nhgis.org/data-availability.

*Census tract and block group boundaries derived from the 2009 TIGER/Line files are available, but NHGIS identifies these boundaries with 2000, not 2009, because they do not completely correspond to the units used in 2009 ACS tables.

TABLE 4.3
NHGIS Crosswalk Availability

Source Zones	Target Zones	1990–2010	2000–2010	2010–2020	2020–2010
Blocks	Blocks	X	X	X	X
Block group parts	Block groups	X	X		
Block group parts	Census tracts	X	X		
Block group parts	Counties	X	X		
Block groups	Block groups			X	X
Block groups	Census tracts			X	X
Block groups	Counties			X	X

Source: https://www.nhgis.org/data-availability.

to the overall housing costs, which the ACS captures both as a percentage of household income or by building age. The latter shifts emphasis from people to housing characteristics such as the number of bedrooms, plumbing, or heating as given in Table 4.3 in the appendix. Table 4.2 in the appendix provides an overview of the tenure variables.

The ACS records estimated selling prices for housing units under "value". It is important to remember that the values are calculated from self-reported estimates of occupied units and vacant units on the market. Housing statistics often suffer from the mixing of different data sources when it comes to such value estimates. As long as one stays with one data source (such as the ACS), data are comparable across time and geography. But ACS values should not be mixed with tax assessments, or the values calculated by licensed assessors for the purpose of securing a loan. Therefore, estimates may become less reliable in a fluctuating or falling housing market. ACS table B25081 (Mortgage Status) records the type of mortgage on owner-occupied housing units. This includes first mortgages, second mortgages, and home equity loans. A simpler classification, "with/without", is used in tables relating mortgage status to topics like real estate taxes, household income, and tenure.

Homeowners with and without mortgages have ongoing monthly costs, and the American Community Survey gathers data about these costs, which are reported in tables referring to "selected monthly owner costs". The costs are reported as either a percentage of the household income or the number of housing units in a monthly cost range such as "$1,250 to $1,499". The selected costs used for these estimates are:

- payments for mortgages, or other debts on the property
- real estate taxes
- fire, hazard, and flood insurance
- utilities (electricity, gas, and water and sewer)
- fuel (oil, coal, kerosene, wood, etc.)
- monthly condominium fees (when applicable)
- mobile home costs (when applicable)

There are two main categories for rent, contract, and gross. Contract rent is the monthly rent agreed to without adjustments for utilities or other payments. Gross rent is similar to selected monthly owner costs. It is the sum of contract rent and the average cost of the utilities (electricity, gas, and water and sewer) and fuels (oil, coal, kerosene, wood, etc.).

The ACS records the number of bedrooms for each housing unit and provides tables to relate that number to tenure and rent. Housing units with only one room are listed as having no bedrooms. The lack of complete facilities for housing units is recorded in two areas: kitchen and plumbing. A complete kitchen requires:

- a sink with a faucet
- a stove
- a refrigerator

Complete plumbing requires:

- hot and cold running water
- a bathtub or shower

If a housing unit doesn’t have one of those items, it is recorded as lacking a complete kitchen or complete plumbing facilities.

Under the rubric of “Selected Conditions”, the ACS describes substandard housing such as:

- incomplete plumbing or kitchens
- overcrowding
- 30% or more of the household income spent on rent or monthly owner costs

4.1.3 National Historical GIS

The US Census Bureau has been collecting data for over 200 years but only post 2000 data can be accessed through their website. To fill this gap, the National Science Foundation funded a long-term project called the National Historical GIS (NHGIS) which is housed at the University of Minnesota. It provides free online access to summary statistics and GIS files for US censuses and other nationwide surveys from 1790 through the present

- County and state census tables since 1790
- Census tract tables since 1910
- Tables for all original census summary levels, down to census blocks, since 1970
- Five-year periods ACS data from 2005–2009 through 2016–2020
- One-year periods ACS data from 2010 through 2019

While all of this is already impressive, NHGIS has also created a plethora of time-series tables that cover a range of basic 100%-count statistics from the 1970 to 2020 censuses as well as several popular sample-based statistics from the 1970 to 2000 long-form surveys and from ACS 5-Year Summary Files for 2008–2012 and 2015–2019. There are also tables of state and county data that go back to 1790 for Total Population and back to 1820 for Persons by Sex. Nominally integrated time series tables, which align geographic units across time by matching names and codes without regard to boundary changes, cover up to eight geographic levels ranging from the nation down to census tracts, see Table 4.2 for more information. The set of covered levels varies among tables according to which statistics are available for each level in each source year. Geographically standardized time series tables that provide estimates for a single year's geographic units by interpolating data from other years, cover 1990, 2000, 2010, and 2020 100%-count statistics for 2010 geographic units at 10 geographic levels ranging from states down to block groups.

The Census Bureau's criteria for the delineation of area units are population-based, e.g., a census tract is supposed to have appr. 4,000 residents – regardless of whether it is in New York or Wyoming. As populations grow (or shrink), the boundaries of the Census area units change, which makes it hard to compare them across years. One of the great features of NHGIS is that they provide crosswalks, i.e., definitions of area units that are consistent across the years. The extent of coverage varies among geographic units and across years. For example, census tracts covered only eight cities in 1910 and did not cover the entire United States until 1990. Table 4.3 provides more detailed coverage information. The basis for NHGIS boundaries before 2000 are 2000 boundary files. For post-2000 boundaries, it is advisable to use 2008 boundary delineations to maintain consistency across the years.

In Chapter 5, we will give examples for how to conduct analyses across years. Very few variables (such as total population) have been consistently measured across the years. Most variable definitions have been undergoing significant changes and the next section will deal with issues of categorical redefinitions. But before we get there, one final but crucial aspect of the US Census data needs to be discussed: the difference between race and ethnicity.

4.1.4 Race and Ethnicity in the US Census Data

The Census Bureau defines race as a person's self-identification with one or more social groups. An individual can report as White, Black or African American, Asian, American Indian and Alaska Native, Native Hawaiian and Other Pacific Islander, or some other race. As of 2000, survey respondents may report multiple races. Ethnicity determines whether a person is of Hispanic origin or not. For this reason, ethnicity is broken out in two categories, Hispanic or Latino and Not Hispanic or Latino. Hispanics may report as any race.

This has multiple confusing consequences. One is that if one adds up all racial observables, the total is larger than the total of the population because an individual may be counted multiple times in different categories. Second, ethnicity is not a racial category and should not be mingled with race counts, see Figure 4.3 which highlights this challenge. Many government statistics do not acknowledge the difference between the two variables, and we often see Hispanic being treated as a racial category. This is wrong and automatically results in faulty statistics. If one is careful, then one can use the Census tables that list the racial categories under Hispanic and

Not Hispanic and create graphics and maps that list each race with its respective ethnic subdivisions – but this is rarely done and still does not solve the issue of multi-racial self-identification. In this volume, when we use racial categories, we limit ourselves to single race declarations only. In the United States-wide context, this is an acceptable generalization, even if it ignores the approximately 10% of the population who in 2020 declared themselves multi-racial. For detailed studies, researchers have to decide whether it is acceptable to follow this approach or whether the inclusion of multi-racial counts paints a better picture of the specific situation.

4.2 FROM MEASUREMENT TO INDICATORS

Primary data collection starts with measurements, using manual or automated counts or by conducting surveys and trusting that people answer honestly. Each measure has a unit of measurement and an expected range of values (for example, the measurement of the number of people in a specific place may increase or decrease but will never include values below zero). The sum of all observations of a measure is a variable. Most housing datasets combine a multitude of variables into tables. In the case of geospatial datasets, one or more of the variables are a spatial reference that associates a record with a specific location.

4.2.1 Locational References

Locational references may come in many shapes and forms. They may be *x*, *y* or latitude/longitude coordinates, addresses, or pointers to well-defined areas such as ZIP code areas, census area units, school districts, etc. The notion of pointers suggests a division of labor, where the details of the locational reference (e.g., the coordinates that make up the boundaries of an area) are stored in one file and the actual measures (house prices, income, etc.) are stored in another file. The pointer then acts as the unique common link between the observation and the location of the observation. This method, known as the geo-relational principle, is quite common with geospatial data as it allows linking multiple datasets to the same location rather than having to store the geospatial details in every dataset (Albrecht, 2007).

Locational references are usually strings – even if the strings consist of a sequence of numbers, which confuses not just users but also many software packages reading geospatial data. ZIP codes are a widely known representative of such numeral strings, where the position of a digit has a hierarchical meaning. For example, ZIP code areas in the US Northeast start with a zero (emphasizing the fact that these are strings rather than numbers), whereas ZIP code areas on the West Coast start with the digit 9. The American National Standards Institute (ANSI) is responsible for maintaining Federal Information Processing Series (FIPS) codes and Geographic Names Information System (GNIS) codes. A wide audience uses FIPS codes and GNIS codes across many private and public datasets to uniquely identify geographic features.

This becomes a lot more important when dealing with the locational references in US Census data, where the GeoID is a fairly long string which is built up from left to right following the schema laid out in Table 4.4. They uniquely identify all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data. From Alaska, the largest state, to the smallest census block in New

TABLE 4.4
The Structure of a US Census GeoID

Area Type	GEOID Structure	Number of Digits	Example GEOID
State	STATE	2	48
County	STATE + COUNTY	2 + 3 = 5	48,201
County subdivision	STATE + COUNTY + COUSUB	2 + 3 + 5 = 10	4,820,192,975
Places	STATE + PLACE	2 + 5 = 7	4,835,000
Census tract	STATE + COUNTY + TRACT	2 + 3 + 6 = 11	48,201,223,100
Block group	STATE + COUNTY + TRACT + BLOCK GROUP	2 + 3 + 6 + 1 = 12	482,012,231,001
Block	STATE + COUNTY + TRACT + BLOCK	2 + 3 + 6 + 4 = 15	482,012,231,001,050

York City, every geographic area has a unique GeoID. Some of the most common administrative/legal and statistical geographic entities with unique GEOIDs include states, counties, congressional districts, core based statistical areas (metropolitan and micropolitan areas), census tracts, block groups and census blocks.

The US Census Bureau uses FIPS codes which are assigned alphabetically by geographic name for states, counties, core based statistical areas, places, county subdivisions, consolidated cities and all types of American Indian, Alaska Native, and Native Hawaiian (AIANNH) areas. Lists of geographic FIPS codes in census products can be found on the ANSI/FIPS Codes page. FIPS codes for smaller geographic entities are usually unique within larger geographic entities. For example, FIPS state codes are unique within the nation and FIPS county codes are unique within a state. Since counties nest within states, a full county FIPS code identifies both the state and the nesting county. For example, there are 49 counties in the 50 states ending in the digits “001”. To make these county FIPS codes unique, the state FIPS codes are added to the front of each county (01001, 02001, 04001, etc.), where the first two digits refer to the state the county is in and the last three digits refer specifically to the county.

The US Census Bureau creates and maintains geographic codes for many statistical geographic areas that are not covered by FIPS codes. These geographic areas include census divisions, census regions, census tracts, block groups, census blocks, and urban areas. The full GEOID for many levels of geography combines both the FIPS codes and Census Bureau codes. For example, census tracts, block groups, and census blocks nest within state and county; therefore, the GEOIDs for each of these geographic areas contain both the state and county FIPS codes, in which they nest. Figure 4.4 illustrates the hierarchical relationship of different geographic areas with one another. Whereas Table 4.4 shows the GEOID structure in TIGER/Line Shapefiles[1] for some of the most common legal and statistical geographies, as well as example GEOIDs for different geographic areas.

4.2.2 Derived Housing Variables

A conceptual model is a representation of a system. Conceptual models are often abstractions of things in the real world, whether physical or social. They consist of concepts

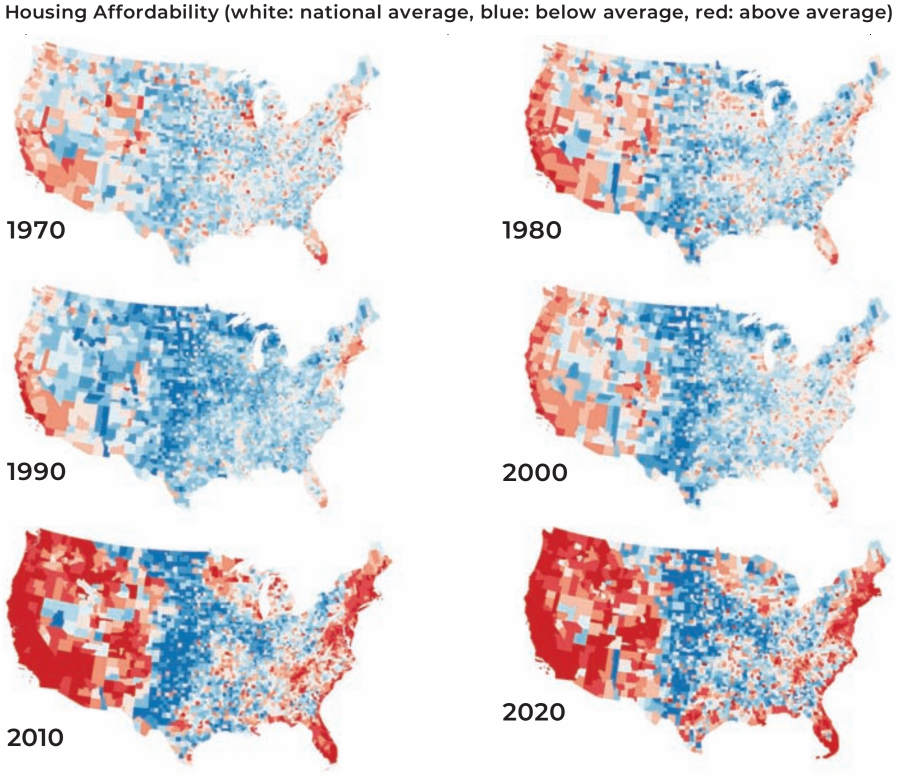

Housing Affordability (white: national average, blue: below average, red: above average)
1970
1980
1990
2000
2010
2020

FIGURE 4.4 Fading affordability: examining housing affordability in the US from 1970 to 2020

used to help people understand the subject the model represents. They are formed after a conceptualization or generalization process. Generalizations posit the existence of a domain or set of elements, as well as one or more common characteristics shared by those elements thus creating a conceptual model. As we acquire and explore data, we will find quite often that in spite of our utmost endeavors, it does not represent our conceptual model. A conceptual model is a representation of a system. Conceptual models are often abstractions of things in the real world, whether physical or social. They consist of concepts used to help people understand the subject the model represents. They are formed after a conceptualization or generalization process. Generalizations posit the existence of a domain or set of elements, as well as one or more common characteristics shared by those elements thus creating a conceptual model.

Typically, the conceptual model of the phenomenon we are investigating requires data that does not exist. If it exists, it may not be aligned elegantly with the conceptual model and therefore require additional manipulation or data wrangling, in industry parlance. The discussion of race and ethnicity in Section 4.1.4 is a good example of the need to transform or map variables from one census year to another. As racial categories were added, we cannot easily compare 1950s or 1970s data with their supposed equivalent from 2000 or later. We may then have to resort to inventing our own variables such as "Non-White" to appropriately represent minority populations.

In other instances, we might want to invent our own indicator for measures that do not exist in the original datasets. Housing affordability is an example of such a derived variable. There are many possible ways to define housing affordability and there are many organizations who have established their own measures. A fairly straightforward measure of housing affordability compares the median rent in an area with the median income in the same area (HUD PD&R Edge, 2017). However, even this simple approach assumes that the majority of people in that area are renters rather than homeowners – which is not the case for the majority of locales in the United States. In that case, we would have to use the current house value, compare it to the area median income, and then weigh this by a measure of how many owners have paid off their mortgage and over how many years the house price should be annualized. Technically, this is all possible, but it illustrates that simple measures can snowball quickly into complicated intractable ones if we want to paint a fair picture across the nation. It is hence the responsibility of the housing policy researcher to be very specific in the definition of their terms and the universe within which they are applicable.

One of the great advantages of working with US Census data is their consistency across the nation. Things get very complicated when we are trying to compare state- or even city-level programs, which in turn tend to have limited life spans, i.e., they are expiring and sometimes replacing each other. A good example of that is the loss of rent-regulated apartments in New York City, which we describe in the following section.

4.3 CHANGE OVER TIME

Although we have espoused the significance and value of understanding the spatial components of housing data, a full understanding of the phenomenon can only be gained if we conceptualize housing and neighborhood change as a spatially differentiated *process*. This requires at a minimum two timestamps for each location and ideally a lot more to capture, for example, the differential aspects of demographic, climate, economic, or policy changes across the country. Everything we observed so far in this chapter still applies but is now compounded by trying to (i) find and (ii) align data across the years. Census tract-level data is exhaustively available only since 1990 and it is hard to imagine these days how little data was collected during the last century overall, and how little of that has been properly archived and curated to be accessible today. Although we are now able to access scanned copies of the NY Times over the last hundred years, many local newspapers have ceased to exist taking their archives (of house prices, for instance) with them. The best source for historical data (with history being as recent as the 1990s) is therefore again the National Historical GIS (NHGIS). As before with the race categories, we need to be conscious of the changing definitions of the variables recorded. We will illustrate this using the example of rent/income changes over time in the following section.

4.3.1 Rent/Income Changes

Among the few variables that have been "consistently" collected over many decades are housing rents and income. It therefore stands to reason that we should be able to study whether housing has become more or less affordable over the years, how different the picture is in different parts of the country, and whether there are any correlations with potentially explanatory variables. Figure 4.5 provides the answer to those questions. In Section 4.9 we outline the conceptual and practical steps it takes to arrive at this figure. The six maps show clear instances of spatial autocorrelation, where likewise values are near each other rather than being randomly distributed within the study area. From a spatial analysis perspective, it would then be interesting to determine the temporal correlation and where socio-economic developments become seeds for the spatial spread of housing affordability (or the lack of) in later years. But this goes beyond the scope of this chapter.

One of the difficulties in working with Census data (despite the US Census being the most consistent and well-documented source of housing data) is that the definitions of variables change over the years. In different Census years, income is accounted for either on a per household or per family basis – and sometimes both. If nothing else is specified, then income is salaried income, excluding transfer payments as well as income from interest, stock options, etc. Tabulations for income have multiplied since the 1970s, when just about the only figure was the median income per area unit. Since then, a variety of other income-related variables have become available, e.g., social security income, aggregate income, the number of people in a particular income bracket, and so on. Similarly, rent started out as just the median rent per area unit but in later years is provided as the number of households in a particular rent bracket (which given the inflationary nature of the subject is changing from decade to decade).

4.4 THE AMERICAN HOUSING SURVEY

The American Housing Survey (AHS) is sponsored by the US Department of Housing and Urban Development (HUD) and conducted by the US Census Bureau. The survey provides information on the size, composition, and quality of the housing across the nation and in major metropolitan areas and measures changes in the housing stock as it ages. The AHS is a longitudinal housing unit survey conducted biennially since 1989 in odd-numbered years. While national data are always collected, typically no more than 30 metropolitan areas are sampled in one survey year. The survey includes questions about:

- the physical condition of homes and neighborhoods,
- the costs of financing and maintaining homes, and
- the characteristics of people who live in these homes.

Planners, policy makers, and community stakeholders use the results of the AHS to assess the housing needs of communities and the country. These statistics inform decisions that affect the housing opportunities for people of all income levels, ages, and racial and ethnic groups. Since the United States changes rapidly, policymakers in government and private organizations need current housing information to make decisions about programs that will affect people of all income levels, ages, and racial and ethnic groups.

HUD uses the AHS to create a biennial Worst Case Needs report to Congress, improve the efficiency and effectiveness of housing programs, and design programs appropriate for different target groups, such as low-income families, first-time home buyers, and the elderly. HUD also uses the data to allocate funds to resolve housing problems, determine qualifications for low-income housing assistance programs, and plan community development (e.g., roads and schools). Academic researchers and

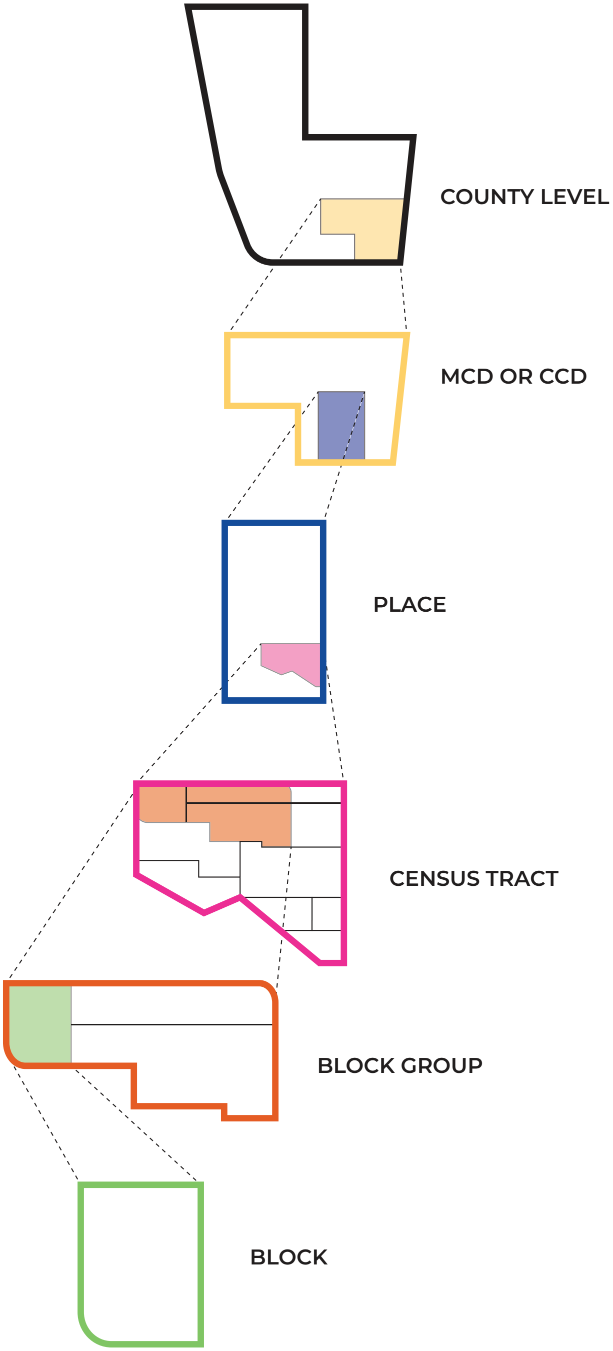


FIGURE 4.5 Unraveling the Census Geography: exploring the interconnectedness of the US Census Bureau's geographic entities and their nested relationships

private organizations also use AHS data to analyze trends in the housing market in efforts of specific interest and concern to their respective communities.

Congress requires the Department of Housing and Urban Development to collect this information under the Housing and Urban-Rural Recovery Act of 1983 (Title 12 of the U.S.C., Section 1701z-1, 1701z-2(g), and 1701z-10a).

Beginning with the 2011 AHS, the survey instrument consists of a permanent core questionnaire plus topical supplements that will rotate in and out of the questionnaire on a yet to be determined schedule. The AHS provides current information on a wide range of "core" housing subjects, including but not limited to the following:

- size and composition of the nation's housing inventory
- vacancies,
- owners and renters,
- physical conditions of housing units,
- equipment breakdowns,
- characteristics of occupants,
- housing and neighborhood quality,
- mortgages and other housing costs,
- fuel usage,
- home improvements,
- persons eligible for and beneficiaries of assisted housing,
- characteristics of recent movers, and
- home values.

In addition to the "core" data, the AHS collected "topical" or supplemental data using a series of modules that will rotate in and out of future surveys. The 2019 topics included:

- home accessibility,
- food security, and
- post-secondary education.

The 2021 AHS includes a mortgage module redesign and the following topical contents:

- Wildfire Risk
- Household Pets
- Secondhand Smoke
- Housing Search
- Intent to Move
- Delinquent Payments and Notices

The 2015 American Housing Survey underwent a major redesign – a new sample was redrawn for the first time since 1985 and new households were asked to participate in the survey, the questionnaire was redesigned, variables were dropped, added, or modified, recodes and imputation methods were streamlined, and the weighting methodology changed. As a result, tables were redesigned, and some estimates became incomparable with previous years.

4.5 ESTABLISHING A GIS DATABASE FOR HOUSING PLANNING RESEARCH

GIS is commonly associated with visualization or more specifically beautifully rendered maps. What is often underappreciated is the fact that GIS relies on large and often complicated databases that reflect the complexity of geographic contexts. Whereas traditional housing research uses one dataset or the other and then represents them in the form of some business graphics, the "I" in GIS is about the (spatial) relationships between different data. The data often comes from different providers, has originally been compiled for different purposes, and in addition to the recoding covered in Section 4.2, now needs to be related to each other. Regardless of whether we are looking at metropolitan or national datasets, the resulting databases often go beyond what can be easily handled on personal computers. In any case, dedicated databases have to be created that should reflect the housing researcher's conceptual model. The common procedure to accomplish this is to build a database schema that captures all the characteristics needed – but no more. Building the final database is as much about removing unwanted variables as it is combining those we seek.

It is beyond the scope of this volume to discuss the foundations of relational database management; suffice it to state here that all the aspects of our research question need to be represented in a collection of tables that are unambiguously linked to each other. Larger organizations will do this in the form of a commercial or open-source database management system like Oracle or Postgres. But every housing researcher is encouraged to mirror the process even in smaller projects by organizing their data in personal database structures such as a geopackage or SpatiaLite.[2]

The first step in organizing one's data is to develop a conceptual model of one's research question. The most common representation of such a conceptual model is a mind map; a listing of all the important aspects of the research question and the relationships between them. If our topic, for instance, is housing insecurity, then we would want to include types of housing insecurity (overcrowding, unsafe housing conditions, eviction, and housing discrimination), factors (poverty, unemployment, rising housing costs, and lack of affordable housing), consequences (foreclosures, evictions, homelessness, housing displacement, poor health outcomes, and economic hardship), demographic groups disproportionately affected by housing insecurity (e.g., low-income households, people of color, and individuals with disabilities), and potential solutions such as affordable housing initiatives, tenant protections, and homelessness prevention programs, see Figure 4.6 for a graphic representation of this type of mind map. For some of these, we would have to determine what variables could serve as indicators, e.g., the number of times a household moves within a year, housing costs as a percentage of income, utility disconnections, or the physical state of the housing infrastructure.[3] Some of these factors are compound variables as in housing costs, which include rent/mortgage, utilities, and property insurance. For each of these factors, the housing researcher then needs to determine the unit of measurement, the spatial and temporal scale (per month or per year, per household or per county), and the likely range of observed values (for quality control purposes, see also the next section).

We recommend that the housing researcher develops this kind of a conceptual model *before* searching for the data to populate their database. There are multiple reasons for that. One is to focus one's mind on the essentials. The resulting database should contain only what we need rather than be the results of an indiscriminate data hunting and

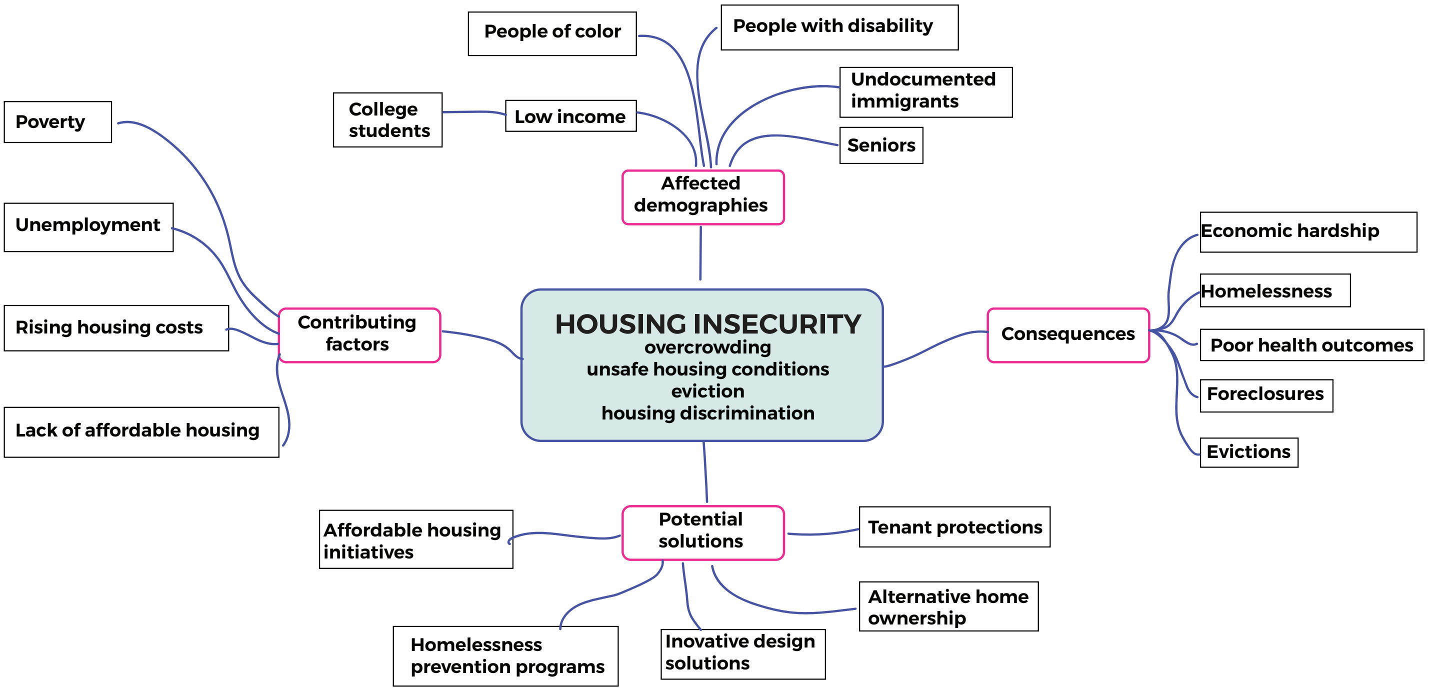


FIGURE 4.6 Mapping housing insecurity: a comprehensive mind map unveiling the causes, effects, affected individuals, and potential solutions to address the crisis

gathering endeavor. Another very important reason is that we do not know what will be useful to us if we have not gone through the exercise of developing the conceptual model. The more rigorous our database schema is, the better equipped we are to know what data we need (to look for) and how to substitute or wrangle the data we are actually getting hold of to satisfy our needs. Finally, the difference between the idealized conceptual model and the actually filled database tells us something about how good the basis for our analysis is. Without the prior development of a conceptual model, we would not be able to judge the quality of the data we are actually dealing with.

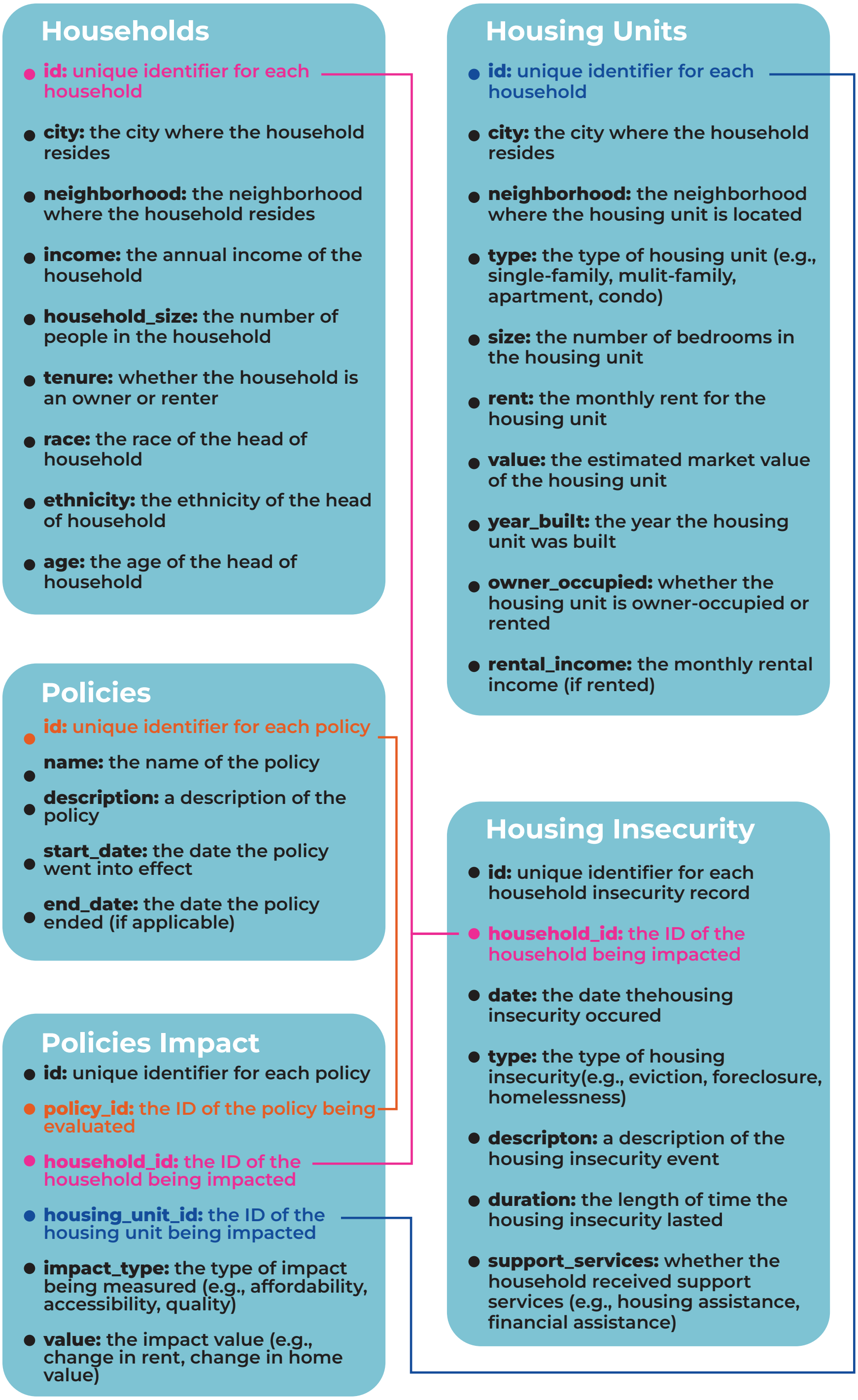
A database schema is the translation of the mind map into an empty database structure. Each of the factors becomes a table for which we have to define what variables it consists of and what datatype is to be used for each variable. Sometimes, this is straightforward as in the setup of household income. Things get a little more complicated when we look at something like the state of the physical infrastructure (doors, windows, walls, roofs, etc.); do we want this measured on a Likert scale and if yes, at what level of aggregation (housing unit, building, census tract)? This is also the time to decide about the spatial reference: do we want the records in our tables to link to an address, an area unit, and an x/y coordinate? The database schema is the well-specified but empty shell of our database. It is defined to exactly fit our needs (which we assessed in the form of our conceptual model). Once our database schema is set up, we are ready to fill the database with data. Sometimes, this is as easy as a one-to-one import of a table into a matching (empty) table in our database. Usually, however, we will select a subset of external tables and have to transform their contents to match the specifications of our database schema. See Figure 4.7 for an example of Housing Insecurity Database Schema.

4.6 DATA QUALITY

As alluded to above, one of the advantages of developing a conceptual model and then designing the database schema accordingly is that any discrepancy between the idealized schema and our adaptations of that schema to match existing data is an indication for how well the data we are working with is suited to truly answer our original research question. Discrepancies between the two are captured by what is known as metadata (data about data). As we seek to fill our own database with data and search the Internet for possible data sources, the metadata tells us how close the external data matches our internal needs. If we cannot find formal descriptions of data, then this sends a warning sign that we might want to be very careful using the data we found.

Official (FGDC- or ISO-conform) metadata consists of many different dimensions of data quality: completeness, spatial, temporal, thematic accuracy, and precision, as well as consistency. In addition, a good metadata documentation will tell us by whom and how the data was generated, for what purpose, how long it is valid, and who is the custodian (from whom we might learn more about it).

Again, the US Census Bureau is the standard against which all other data sources can be measured by. For a novice user of Census data, the exhaustive description of data quality that is directly embedded into the data rather than in a separate metadata document can be stressful. Every ACS variable is accompanied by its respective Margin of Error (MoE) at the 90% confidence level. This implies a 10% chance of incorrect inference for all estimates, see Table 4.5.



Households

- **id:** unique identifier for each household
- **city:** the city where the household resides
- **neighborhood:** the neighborhood where the household resides
- **income:** the annual income of the household
- **household_size:** the number of people in the household
- **tenure:** whether the household is an owner or renter
- **race:** the race of the head of household
- **ethnicity:** the ethnicity of the head of household
- **age:** the age of the head of household

Housing Units

- **id:** unique identifier for each household
- **city:** the city where the household resides
- **neighborhood:** the neighborhood where the housing unit is located
- **type:** the type of housing unit (e.g., single-family, mulit-family, apartment, condo)
- **size:** the number of bedrooms in the housing unit
- **rent:** the monthly rent for the housing unit
- **value:** the estimated market value of the housing unit
- **year_built:** the year the housing unit was built
- **owner_occupied:** whether the housing unit is owner-occupied or rented
- **rental_income:** the monthly rental income (if rented)

Policies

- **id:** unique identifier for each policy
- **name:** the name of the policy
- **description:** a description of the policy
- **start_date:** the date the policy went into effect
- **end_date:** the date the policy ended (if applicable)

Housing Insecurity

- **id:** unique identifier for each household insecurity record
- **household_id:** the ID of the household being impacted
- **date:** the date thehousing insecurity occured
- **type:** the type of housing insecurity(e.g., eviction, foreclosure, homelessness)
- **descripton:** a description of the housing insecurity event
- **duration:** the length of time the housing insecurity lasted
- **support_services:** whether the household received support services (e.g., housing assistance, financial assistance)

Policies Impact

- **id:** unique identifier for each policy
- **policy_id:** the ID of the policy being evaluated
- **household_id:** the ID of the household being impacted
- **housing_unit_id:** the ID of the housing unit being impacted
- **impact_type:** the type of impact being measured (e.g., affordability, accessibility, quality)
- **value:** the impact value (e.g., change in rent, change in home value)

FIGURE 4.7 A housing insecurity database schema based on the mind map of Figure 4.6

TABLE 4.5
Sample US Census Table With the Margin of Error Information

Tract	Total	Utility Gas	Bottled or LP Gas	Electricity	Fuel Oil	Coal	Wood	Solar	Other Fuel	No Fuel Used
1.01										
Estimate	2,226	1,707	0	490	17	0	0	0	0	12
MoE	±302	±334	±9	±173	±27	±9	±9	±9	±9	±18
1.02										
Estimate	1,591	847	0	699	0	0	0	0	13	32
MoE	±151	±133	±9	±126	±9	±9	±9	±9	±19	±32
2.01										
Estimate	1,747	1,106	9	607	25	0	0	0	0	0
MoE	±184	±205	±14	±178	±35	±13	±13	±13	±13	±13
2.02										
Estimate	2,415	1,764	0	520	0	11	120	0	0	0
MoE	±266	±242	±13	±175	±13	±19	±188	±13	±13	±13
3.01										
Estimate	622	507	0	61	9	0	6	0	18	21
MoE	±44	±52	±9	±33	±9	±9	±8	±9	±14	±13

Most other data sources will not have such intricate quality information down to the individual record level. But at a minimum, there should be a separate metadata document (preferably following an established standard such as *Dublin Core* or *FGDC*) and a data dictionary.[4] The lack of such documentation suggests poor data quality in the first place and hence limited reliability for our data analysis down the road.

Most other data sources will not have such intricate quality information down to the individual record level. But at a minimum, there should be a separate metadata document (preferably following an established standard such as *Dublin Core* or *FGDC*) and a data dictionary.[5] The lack of such documentation suggests poor data quality in the first place and hence limited reliability for our data analysis down the road.

4.6.1 Data-Poor Environments

As soon as we move beyond federal data collection efforts, we will find that housing data is getting sparse. Few states and even fewer municipalities or non-profit organizations have the resources to collect housing-related data. Companies (especially utilities) are not prone to share their data and the data collected by academic organizations tends to be limited in spatial and temporal scope. The result is a patchwork of data that is impossible to generalize. In Section 4.8, we will identify a few non-conventional data sources but in the meantime, the onus is on the individual researcher to peruse data portals such as the *National Neighborhood Data Archive* at the University of Michigan, the UC Berkeley's *Urban Displacement Project*, Esri's *ArcGIS Data Hub*, the U.*S. city open data census*, or general purpose repositories such as *Awesome Public Datasets* or *Kaggle*.

4.7 SCALE ISSUES

We have emphasized that GIS-based housing research relies heavily on combining different datasets. The moment we do this, however, there is a good chance that the datasets have been compiled at different scales/resolutions, e.g., counties vs. metropolitan areas, or ZIP code areas vs. census tracts. One of the main functions of GIS is to overlay and disaggregate the respective datasets to create area units with aligned boundaries. We will discuss those techniques in detail in the following chapter; however, regardless of how we proceed, there are a few methodological issues that housing researchers need to be aware of, the most prominent among them being the modifiable area unit problem (see the following sub-section).

Even if we do not combine different datasets, many housing-related datasets are either spatially incomplete (say, they cover only urban areas in the United States but exclude rural ones), or they have highly varying spatial footprints within one and the same dataset. An example of the latter is HUD Continuum-of-Care program data, see Figure 4.8 (USHUD, n.d.), which sometimes are as big as a whole state and sometimes as small as a mid-sized city (e.g., Fall River, MA).

Other datasets, such as the US EIA's 2022 residential energy consumption survey, initially look impressive as it is based on 18,496 survey respondents – but of its over 100 variables, the finest spatial resolution is that of a state. If we then look at a variable such as the frequency of disconnection notices, we find that this survey is not at all representative, as not a single survey respondent has received such a notice more than once a year – which is inconsistent when compared to the American Housing Survey (2013),6 according to which some 7.8% of all households surveyed received such notices.

4.7.1 The Modifiable Area Unit Problem (MAUP)

The modifiable area unit problem (MAUP) is a summary term for two different but related issues when dealing with spatial data. The first is the issue of scale and this is what statisticians call ecological fallacy or the fact that we cannot draw conclusions about specifics from the aggregate. If we know what percentage of the vote a presidential candidate received in a state, then this tells us nothing about how they performed in one of the state's counties. The other aspect of MAUP is unique to spatial data and unfortunately very common. The name "modifiable area" points to the issue of different possible ways to subdivide an area. We will demonstrate this by looking at the various ways different city agencies are carving up New York City.

Figure 4.9 (NYC Open Data, n.d.) illustrates the boundary problem. In a typical housing GIS project, we would compile data from different city agencies. Many of these have their own administrative boundaries; while there are many more, this example shows borough/county boundaries (the city of NYC encompasses five counties), community districts, neighborhood planning areas as defined by the Mayor's office, police precincts, postal ZIP code area boundaries, inclusionary housing areas where zoning has been restricted to support affordable housing, and finally the boundaries around different kinds of zoning (related to but not the same as inclusionary housing). These do not even include school districts or public health planning areas nor many of the special areas such as flood zones, etc.

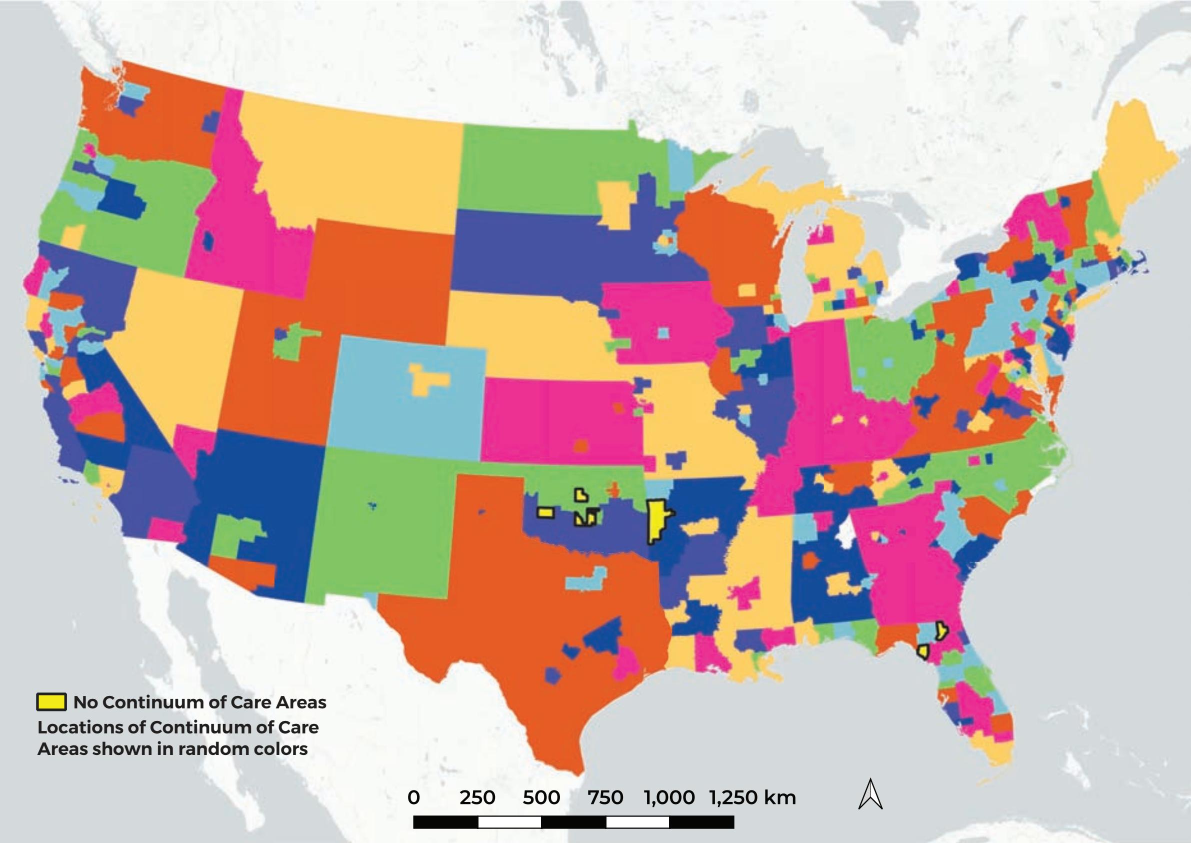

🟨 No Continuum of Care Areas
Locations of Continuum of Care Areas shown in random colors
0 250 500 750 1,000 1,250 km

FIGURE 4.8 HUD continuum-of-care area units

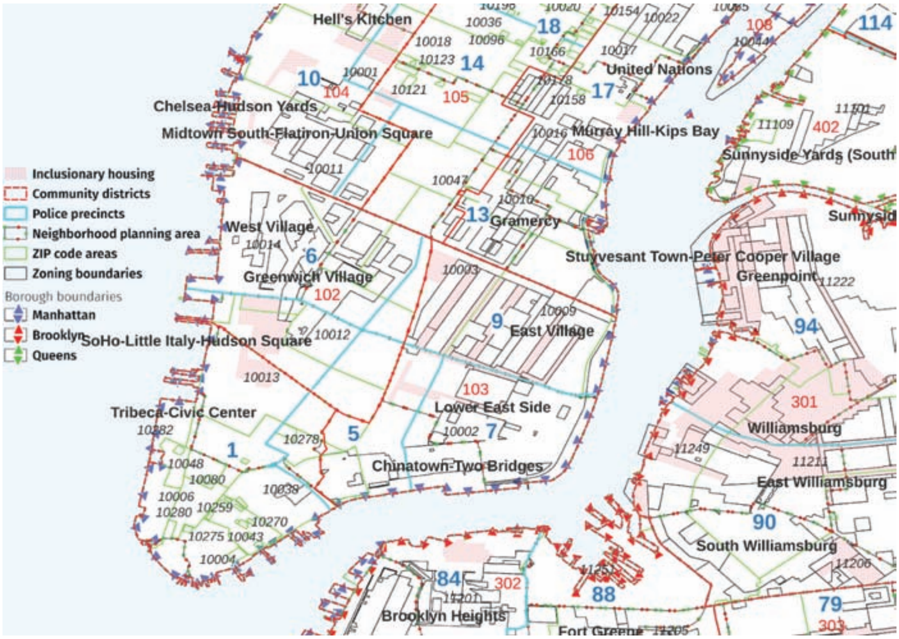


FIGURE 4.9 Sample snapshot of different administrative boundaries in New York City

Underlying each of these different boundaries are GIS layers that contain a range of descriptors (attributes), which we might want to make use of in our comprehensive analysis. The question then arises, what would happen if any of these city agencies had drawn their boundaries differently? Regardless of what area unit we are looking at, it is the aggregation of individual events (e.g., postal addresses, crime locations, and gerrymandered political boundaries) that then results in aggregate values that *are a function of* how the boundaries are drawn. In other words, if the boundaries have been drawn in a different fashion, the observed aggregate values would be different and would play a different role in our analysis. Research (Openshaw and Taylor, 1979) has shown that, to take an extreme political example, it is possible to redraw electoral boundaries in such a way that in almost every state 100% of all representatives hail from one party only.

This problem would not occur if all our areal boundaries would coincide; if for instance, the postal, electoral, planning, and administrative boundaries (police, school, etc.) would all either coincide or neatly place into each other as many of the US Census boundaries do. The New York City department of City Planning is spending considerable efforts trying to align at least the zoning and planning-related boundaries with those of the Census Bureau in an attempt to minimize the effects of the MAUP. Way ahead of us is France, where most official boundaries align in a neat hierarchical fashion similar to what we discussed about the spine in the US Census hierarchy of area units – except that in France, this applies to postcodes, car license plates, fire and school districts, etc.

4.8 NON-CONVENTIONAL DATA SOURCES

In Section 4.6, we described the dearth of data in many aspects of housing research. In this section, we are going to point to a few data sources that are not in the realm of official data but may yet be quite useful.

4.8.1 Real Estate Brokerages and Consolidators

The first few are actually quite obvious; as the housing sector (in the United States) is dominated by private businesses, they have an interest in collecting relevant data. Usually, such datasets are proprietary, and many companies seek to maintain their competitive advantage by not disclosing their data. But there are a few exceptions. Redfin is one of the big brokerage companies and was the first one to use a simple online GIS to advertise their properties. They release *weekly, monthly, and quarterly datasets* with several million records each at a spatial resolution of counties and/or metro areas. There are a lot of redundancies in these datasets that require a bit of data wrangling, and the housing researcher would also have to compile their own set of geometries for the counties and metro areas to eventually perform a spatial join (see Chapter 5) to incorporate these datasets into GIS.

Zillow (including its merger with Trulia) and Realtor.com are meta websites that serve real estate agents but also allow individual sellers to list their properties. Their business is to compile the non-standardized records of multiple listing services (MLS) from around the country, resulting in a very comprehensive overview of the residential real estate market. Zillow has both *data download options* as well as an application programmers' interface (API). In addition to owned property data, Zillow also publishes monthly *rental* rates on a per ZIP code areas basis going back to 2014. The Zillow *API allows* developers to query their vast database down to the individual property level, which includes property tax information for almost 150 million properties in the United States. One of their most widely used datasets is a delineation of some *17,000 neighborhood boundaries* in 650 cities which is now hosted by the US EPA.

4.8.2 HAZUS Housing Stock Data

A less obvious source of housing data is the Department of Homeland Security's *HAZUS MH program*. An add-on to ArcGIS Desktop, HAZUS is used to model the physical, economic, and social impacts of disasters. The software is of limited use to housing researchers; however, the program comes with extensive datasets needed to estimate potential losses derived from *Homeland Infrastructure Foundation-Level Data* (HIFLD) and the National Structure Inventory (NSI) produced by the US Army Corps of Engineers. The NSI does provide a structure-level representation (as points) of most structures in the U.S., as well as multiple building characteristics including type, occupancy, construction date, building material, utilities connected, etc. However, it is far from perfect and, as with every dataset, should be carefully evaluated to determine if it is suitable for the purpose of any given study.

4.8.3 Individual Data Collection; Windshield Survey; Crowdsourced Data

If none of the resources discussed in this chapter fulfill the needs of the housing researcher, then the last resort is to embark on one's own data collection. Sometimes, this is as straightforward as conducting a walking or windshield survey to examine more specific facets of a neighborhood such as

- The age, nature, and condition of the community's available housing
- Infrastructure needs – roads, bridges, streetlights, etc.
- The presence or absence of functioning businesses and industrial facilities
- The location, condition, and use of public spaces
- The amount of activity on the streets at various times of the day, week, or year
- The amount and movement of traffic at various times of day

Windshield surveys require "boots-on-the-ground" but can be a very efficient way of data gathering – especially in a participatory research context. Neighborhood-based researchers can rapidly compile a list of desirable and objectionable characteristics, especially if equipped with mobile phone-based location recording software such as *KoBoToolbox* or *Survey123*. The advantages are the same as for any primary data collection: with complete control over the survey design and collection process, the appropriateness of the data is guaranteed. And if the data collection is performed by locals for locals, then a certain degree of buy-in can be assumed, which helps with respect to quality control. Numerous studies have shown that crowd-sourced geospatial data such as Open Street Map (OSM) is equivalent and sometimes even superior to authoritative data (Zielstra and Zipf, 2010; Zhang and Pfoser, 2019; Jacobs and Mitchell, 2020), which caused, for example, the New York City government to create an agreement with OSM to regularly exchange updates to their respective databases, resulting in one of the best municipal datasets world-wide.

The obvious disadvantage is that the data collection effort will be limited in spatial and temporal scope and cannot be generalized beyond the neighborhood or small city level. Larger surveys become prohibitively expensive, even for experienced and well-funded organizations. The best way to collect housing data on a national scale is to attach the data collection effort to the work of a larger volunteer organization such as the National Low Income Housing Coalition, National Fair Housing Alliance, the National Association of Housing and Redevelopment Officials (NAHRO), the Council for Affordable and Rural Housing, the National Association of Housing Cooperatives, or the National Civic League.

4.9 GIS ACTIVITY

In Section 4.3, we presented a figure that illustrates how Housing Affordability has changed over the years and how it also changes geographically. In this section, we are describing the steps it takes to arrive at the maps of Figure 4.4. The topic of this figure is the notion of housing affordability and how it is expressed differently in different parts of the county and developed over time. This starts with a definition of housing affordability. When you google for this term, you will invariably come across the figures of the National Association of Realtors, which by definition covers only (potential) property owners. They release monthly data on a per ZIP code area basis (but have a very restrictive data use policy) that varies mostly because of the month-to-month changes in mortgage interest rates. For the purposes of our example here, we are looking at affordability not just from a homeowner's perspective but every form of tenure.

Our conceptual model takes into consideration rent as a percentage of household income as well as the value of a home in relationship to the owner's income. Neither of these figures are available on a per-household basis. Given that the map in Figure 4.7 covers the whole nation, we decided that county-level data is the appropriate spatial resolution. There are a little over 3,300 counties in the United States, which exhaust the variability that a human observer can handle on a single map. Alternatively, the same data is available at census tract resolution for regional analyses.

Our conceptual model treats renters and homeowners separately. Renters pay their rent on a monthly basis (which is also how it is recorded by the Census Bureau), while homeowners accumulate their assets over the lifetime of their mortgage. Both have additional housing-related expenses such as heating and insurance. But these are complications that do not influence the basic conceptual model. The Census Bureau has been collecting data about mortgage payments but only as of late, making longitudinal analyses impossible. We therefore chose to annualize monthly rents to match annual income values and to spread property values over a 30-year period and then take the annual value as a percentage of the annual income. All the values are using the median values per Census area unit (in our case counties).

For the decadal years 1970–2000, we used data from the NHGIS website, while for 2010 and 2010 we retrieved the raw data from the US Census website. It turns out that for 2010, the US Census Bureau lists ACS 1-year data for only 820 counties, so we had to use the 5-year ACS data for 2010. The universe of counties for the 48 conterminous states varies between 3,008 and 3,011 counties, which has no discernable effect on our maps in Figure 4.4 but constrains a spatio-temporal analysis to only those counties that exist consistently across the five decades.

The ratio of homeowners to renters varies widely across the country. Our calculation of housing affordability therefore weighs the rent burden and homeownership costs according to the percentages of those two categories in each county. After downloading the respective datasets and deriving the base variables for each decade, the calculation of housing affordability is now consistent across the decades. The final step is to calculate the difference in affordability for each county compared to the median value of all counties in the 48 conterminous states. The respective maps depict the difference in shades of red (less affordable than the nationwide median) and blue (more affordable than the nationwide median). The first impression is that housing affordability was much more evenly distributed in the 1970s than in the 2010s. A lot more counties were close to the national median back then than there are now. Affordability was much less an issue in the 1990s than as of late. Particularly striking is the change in the Mountain West where large swaths of the country changed from very affordable to the opposite in only 20 years. The northern Nevada holdout then gave way in 2020 as well.

NOTES

1. TIGER/Line Shapefiles will be explained in Chapter 5.
2. We will discuss geospatial data formats and storage mechanisms in the following chapter.
3. The actual list is clearly a function of the research question at hand and is likely to differ depending on who is asking it.
4. See, for example, https://files.hudexchange.info/resources/documents/FY-2022-HMIS-Data-Dictionary.pdf.
5. See, for example, https://files.hudexchange.info/resources/documents/FY-2022-HMIS-Data-Dictionary.pdf.
6. Unfortunately, the bi-annual AHS has not asked this question since 2013.

FURTHER READING

Chu, M, Fenelon, A, Rodriguez, J, et al., 2022. "Development of a Multidimensional Housing and Environmental Quality Index (HEQI): Application to the American Housing Survey". *Environ Health*, 21: 56. doi:10.1186/s12940-022-00866-8.

Devillers, R, and Jeansoulin, R, 2006. *Fundamentals of Spatial Data Quality.* doi:10.1002/9780470612156.

Donnelly, F, 2022. "US Census Data: Concepts and Applications for Supporting Research". *American Library Association Library Technology Reports,* 58: 4.

FEMA, 2022. *Hazus 6.0 Baseline Data Updates. FEMA Factsheet*. https://www.fema.gov/sites/default/files/documents/fema_hazus-6-data-updates-factsheet.pdf, last accessed 4 December 2022.

Goodchild, M, W, Li, and Tong, D, 2022. "Introduction to the Special Issue on Scale and Spatial Analytics". *Journal of Geographical Systems*, 24: 285–289. doi:10.1007/s10109-022-00391-9.

Guptill, S, and Morrison, J, 1995. *Elements of Spatial Data Quality*. Amsterdam: Elsevier.

Hirschman, C, Alba, R, and Farley, R, 2000. The Meaning and Measurement of Race in the U.S. Census: Glimpses into the Future. *Demography*, 37: 381–393.

Missouri Census Data Center, 2022. *Intro to Census Geography, Summary Levels, and GeoIDs*. https://mcdc.missouri.edu/geography/sumlevs/, last accessed April December 2022.

National Institute of Standards and Technology (NIST), 2021. *Compliance FAQs: Federal Information Processing Standards (FIPS).* https://www.nist.gov/standardsgov/compliance-faqs-federal-information-processing-standards-fips, last accessed 4 December 2022.

Sparx Systems, 2022. *Guide to Business Modeling*. https://sparxsystems.com/resources/user-guides/16.0/guidebooks/business-modeling-techniques.pdf, last accessed 4 December 2022.

Uhl, J, Leyk, S, McShane, C, Braswell, A, Connor, D, and Balk, D, 2021. "Fine-Grained, Spatio-Temporal Datasets Measuring 200 Years of Land Development in the United States". *Earth System Science Data*, 13(1):119–153. doi:10.5194/essd-13-119-2021.

USHUD, n.d. *Continuum of Care GIS Tools. HUD Exchange*. Washington, DC: Department of Housing and Urban Development. Online resource available at https://www.hudexchange.info/programs/coc/gis-tools/, last accessed 12 July 23.

Walker, K, 2023. *Analyzing US Census Data*. London: Chapman & Hall.

Wong, D, 2004. The Modifiable Areal Unit Problem (MAUP). In Janelle, D, Warf, B, and Hansen, K (Eds.), *WorldMinds: Geographical Perspectives on 100 Problems*. Dordrecht: Springer. doi:10.1007/978-1-4020-2352-1_93.

REFERENCES

Albrecht, J, 2007. *Key Concepts and Techniques in GIS*. London: Sage Publications.

Chu, M, Fenelon, A, Rodriguez, J, Zota, and Adamkiewicz, G, 2022. "Development of a Multidimensional Housing and Environmental Quality Index (HEQI): Application to the American Housing Survey". *Environmental Health,* 21: 56. doi:10.1186/s12940-022-00866-8.

Devillers, R, and Jeansoulin, R, 2006. *Fundamentals of Spatial Data Quality*. doi:10.1002/9780470612156.

Donnelly, F, 2022. "US Census Data: Concepts and Applications for Supporting Research". *American Library Association Library Technology Reports,* 58: 4.

FEMA, 2022. "Hazus 6.0 Baseline Data Updates". *FEMA Factsheet*. https://www.fema.gov/sites/default/files/documents/fema_hazus-6-data-updates-factsheet.pdf, last accessed 4 Decmber 2022.

Gold, C, 2016. "Tessellations in GIS: Part I-putting it all together". *Geo-spatial Information Science*, 19(1): 9–25, doi:10.1080/10095020.2016.1146440.

Goodchild, M, Li, W, and Tong, D, 2022. "Introduction to the Special Issue on Scale and Spatial Analytics". *Journal of Geographical Systems*, 24: 285–289. doi:10.1007/s10109-022-00391-9.

Guptill, S, and Morrison, J, 1995. *Elements of Spatial Data Quality*. Amsterdam: Elsevier.

Hirschman, C, Alba, R, and Farley, R, 2000. "The Meaning and Measurement of Race in the U.S. Census: Glimpses into the Future". *Demography*, 37: 381–393.

HUD, and PD&R Edge, 2017. *Defining Housing Affordability*. Washington, DC: U.S. Department of Housing and Urban Development's (HUD's) Office of Policy Development and Research (PD&R). https://www.huduser.gov/portal/pdredge/pdr-edge-featd-article-081417.html, last accessed 4 March 2023.

Jacobs, K, and Mitchell, S, 2020. "OpenStreetMap Quality Assessment Using Unsupervised Machine Learning Methods". *Transactions in GIS*, 24: 1280–1298. doi:10.1111/tgis.12680.

Missouri Census Data Center, 2022. *Intro to Census Geography, Summary Levels, and GeoIDs*. https://mcdc.missouri.edu/geography/sumlevs/, last accessed 12/04/2022.

NYC Open Data, n.d. *NYC Open Datamine*, online resource available at https://opendata.cityofnewyork.us/data/.

Openshaw, S, and Taylor, P, 1979. "A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem". In Wrigley, N (Ed.), *Statistical Methods in the Spatial Sciences*, pp. 127–144. London: Pion.

Ramasubramanian, L, and Albrecht, J, 2018. *Essential Methods for Planning Practitioners: Skills and Techniques for Data Analysis, Visualization, and Communication.* The Urban Book Series. Cham: Springer. doi:10.1007/978-3-319-68041-5.

Sparx Systems, 2022. *Guide to Business Modeling*. https://sparxsystems.com/resources/user-guides/16.0/guidebooks/business-modeling-techniques.pdf, last accessed 4 December 2022.

Uhl, J, Leyk, S, McShane, C, Braswell, A, Connor, D, and Balk, D, 2021. "Fine-grained, spatio-temporal datasets measuring 200 years of land development in the United States". *Earth System Science Data* 13(1):119–153. doi:10.5194/essd-13-119-2021.

US Census, 2007. *Measuring Overcrowding in Housing*. Report written by the U.S. Department of Housing and Urban Development. Online resource, https://www.census.gov/content/dam/Census/programs-surveys/ahs/publications/Measuring_Overcrowding_in_Hsg.pdf, last accessed 4 March 2023.

US Census, 2020. *Census CQR Frequently Asked Questions (FAQs)*. https://www2.census.gov/programs-surveys/decennial/2020/program-management/cqr/cqr-faqs.pdf, last accessed 16 May 2023.

Walker, K, 2023. *Analyzing US Census Data*. London: Chapman & Hall.

Wong, D, 2004. "The Modifiable Areal Unit Problem (MAUP)". In Janelle, D, Warf, B, and Hansen, K (Eds.), *WorldMinds: Geographical Perspectives on 100 Problems*. Dordrecht, NL: Springer. doi:10.1007/978-1-4020-2352-1_93.

Zhang, L, and Pfoser, D, 2019. "Using OpenStreetMap Point-of-Interest Data to Model Urban Change-A Feasibility Study". *PLoS ONE* 14(2): e0212606. doi:10.1371/journal.pone.0212606.

Zielstra, D, and Zipf, A, 2010. "Quantitative Studies on the Data Quality of OpenStreetMap in Germany". In *Proceedings of the Sixth InternationalConference on Geographic Information Science,* pp. 20–26. Zurich, Switzerland: GIScience, University of Zurich.

5 GIS Analysis and Visualization

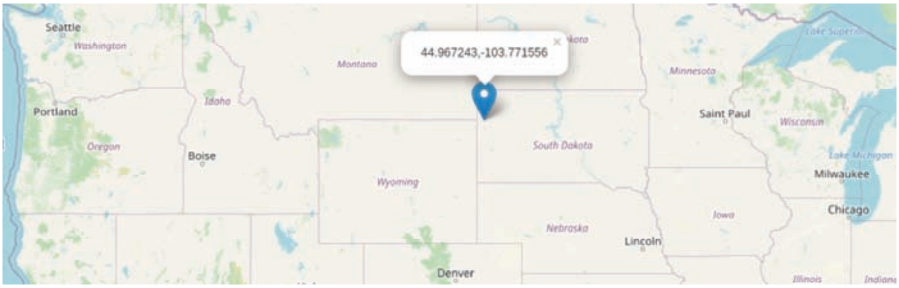
5.1 GIS CORE CONCEPTS

The unique feature that distinguishes GIS from all other software is its ability to help researchers analyze data spatially i.e., to reason about spatial relationships. We have, in the previous chapter, written about the need to organize our data and conveyed our preference for using a database management system (DBMS). Traditional database management systems are not equipped to handle spatial data. End users have to link and integrate spatial information in order to conduct spatial analyses. So, we have to ask, *what is it that makes spatial special*? This chapter cannot replace a formal primer or course about GIS fundamentals. Our goal is instead to provide the thoughtful and serious reader with enough information, so that she can engage in meaningful conversations with GIS specialists. As we discuss the core concepts of GIS, we first have to cover four foundational concepts: (i) coordinate reference systems (CRS), (ii) spatial data types, (iii) spatial operations, and (iv) the geo-relational principle –all of which are unique to GIS, although they can be added to DBMSs. The first three concepts deal with the special spatial nature of GIS-based reasoning and communication; CRS determine where on the surface of this planet our data pertains to, spatial data types are a necessary ingredient to deal with the multi-dimensionality of spatial information, and spatial operations are what allow us to measure distances and directions, and analyze spatial relationships such as adjacency, intersections, or containment. A foundational aspect of GIS is the georelational principle, where every piece of information that we store in a GIS has both a spatial footprint and a set of attributes that describe what we find at the footprint's location.

5.1.1 Coordinate Reference Systems

When we work with spatial data, we are usually describing a location on Earth and attempting to describe what we can observe at that precise location (more about this in Section 5.1.2). The location can be a point (e.g., city), or a line (e.g., street), or an area (e.g., county) that has a unique position on the Earth's surface. Coordinate reference systems are used to describe that position (and if the location is larger than a point also the geometric shape of that location). The tricky thing here is the fact that Earth is a spherical object, and that spherical geometry is (i) really hard and (ii) difficult to communicate. Imagine, if you will, attempting to take the entire peel of a juicy orange fruit (a three-dimensional object) and laying it flat on a table and then trying to link a point on the peel to a point on the peeled orange's surface. Whenever we try to transpose a location on the Earth's surface onto a two-dimensional plane, so that we can apply the geometric rules we learned in middle school, we are compromising one geometric characteristic (size, shape, direction, and distance) or the other. Hundreds of different

DOI: 10.1201/9780429243066-5



Coordinate System	North	West	Zone	Projected
Latitude / longitude	44°58'2.07622"	-103°46'2.07622"		No
Decimal degree	44.967243	-103.771556		No
Universal Transverse Mercator	4,980,045.51	596,875.35	13T	Yes
SPC 1927 (feet)	1,024,338.727	436,137.9431846		Yes
SPC 1983 feet	992,719.62101	436,124.6079814		Yes
SPC 1983 meter	302,580.76024	132,931.8866112		Yes

FIGURE 5.1 Center of the nation: exploring various coordinate values identifying the geographic center of the United States

coordinate systems have been developed to minimize the distortions and they are all incompatible, i.e., if we combine data that is encoded based on different coordinate systems, we have to translate it from one into the other. A GIS (as well as a spatially enabled DBMS) therefore has to incorporate a library of all the different coordinate systems and has to be able to translate data between the different encodings.

Once we have established what coordinate system to use to describe our positions, we have to decide whether to use two, three, or four values (*x*, *y*, *z*, *t*) to encode a point in two or three spatial dimensions, as well as potentially in time (to capture movements or change). Points are then combined into lines, which are combined into areas, and potentially volumes to describe the spatial phenomenon of interest. Depending on the coordinate system used, we deal either with x and y values or with the latitude and longitude values of spherical geometry. As a rule of thumb, if the coordinate values are small (maximum three digits before the decimal point) then we are dealing with spherical coordinates, whereas if the values are large (in the hundreds of thousands or millions), then we are dealing with coordinates that are projected onto a plane.

Figure 5.1 illustrates the havoc created if the particular coordinate system is not specified. All the locations refer to the same exact position on the Earth's surface. If the coordinate system definition is not provided and instead (wrongly) assumed, the center of the United States may jump around between Wyoming, South Dakota, and Canada, and in the extreme case of assigning unprojected coordinates right on the equator.

5.1.2 Spatial Data Types

When we store geographic data, we are either describing features whose locations are given in the form of some geometry or we are describing regular tessellations of space (see Section 5.3.2). In either case, we are dealing with complicated structures that cannot be represented by the data types commonly used in spreadsheet or database programs.

TABLE 5.1
Simple Point Geometries in a CSV File, Stored Together with Attribute Information

XY	Lat	Lon	Address	Function	Capacity	Year	Revenue
6274.97, 428422.31	38.30742	−102.8561	26 Mall Dr., Town, Zip	Mall	4,200	2006	974.2

GIS and spatially enabled databases have special data types that allow to store 2- or 3-dimensional coordinates and then combine these into higher-dimensional geometries of variable length. Spatial data types bear some similarity to temporal data types, where we have many different ways to store data and time. Yet, spatial data types are more complicated because of the need to uniquely reference multiple dimensions. To illustrate the point, imagine we are storing city locations as *X*, *Y* values. *X* and *Y* have to be treated not as separate fields but as a tuple (a singular entity consisting of two parts) because if we treat the North and the West values in Table 5.1 independently, then we could sort locations by their "Northness" regardless of their "Westness".

When we store the coordinates that make up the lines of rivers or the boundaries of counties, then there is no way for us to know in advance how many coordinates we need to encode and store a particular county or river. This means that we need data types that allow for variable length of the values stored in them. Alternatively, if we are dealing with spatial phenomena that have no well-defined boundaries, then we can store their spatial footprint in a data structure that is similar to an image – which is yet another spatial data type.

In general, data types designate the amount of memory used to store the data and the internal organization. In addition, data types determine what kind of operation can be performed on the data stored using one type or the other. For instance, when we store spreadsheet data as type character or date, then we cannot perform multiplications on those values. The same is true for spatial data types: once chosen, we are limited to the kind of operations that are applicable for one (spatial) data type of the other. Regardless of what spatial data type we choose the coordinates that we use to store our spatial data are a function of the previously chosen coordinate reference system.

5.1.3 Spatial Operations

Spatial operations can be coarsely divided into quantitative and qualitative ones. On the quantitative side, we have measurements of distance and direction, as well as subsequent calculations of areas and volumes. On the qualitative side, we have topological relationships such as inside, outside, touching, and intersecting/crossing. To the uninitiated, these may seem to be fanciful but they are essential for spatial reasoning as well as quality control. If we assume (and this is a rather bold assumption) that all our data is encoded using the same projected coordinate reference system, then we could use drawing programs or CAD systems to calculate distance and direction. But the ability to check whether a particular road crosses a river or a city boundary even if the two do not share a common recorded point is one of the hallmarks of GIS software that no other software is equipped to handle, see Figures 5.2 and 5.3.

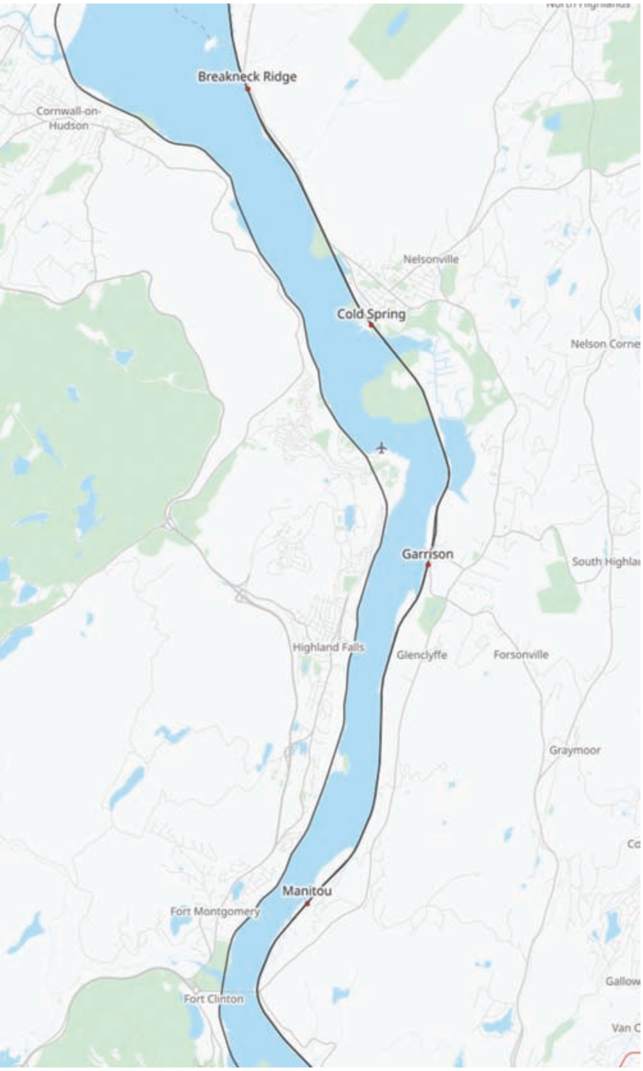


FIGURE 5.2 Is the road crossing the river or staying on one side?

5.1.4 The Georelational Principle

Throughout this volume, we have talked about geographic entities that are a combination of a spatial footprint and the characteristics that we observe at the location of that footprint. The one-to-one relationship between the two is known as the geo-relational principle. It is mirrored after the basic theorem underlying relational databases, where we link records in one table with records in another table. Applied to geographic data, this link is now between a record describing a specific geometry at a unique location and another record in a table of non-spatial attributes. This requires each table to have a field containing a unique value (primary key) and there to be one and only one corresponding record containing the same key values in the two tables that form the relationship.

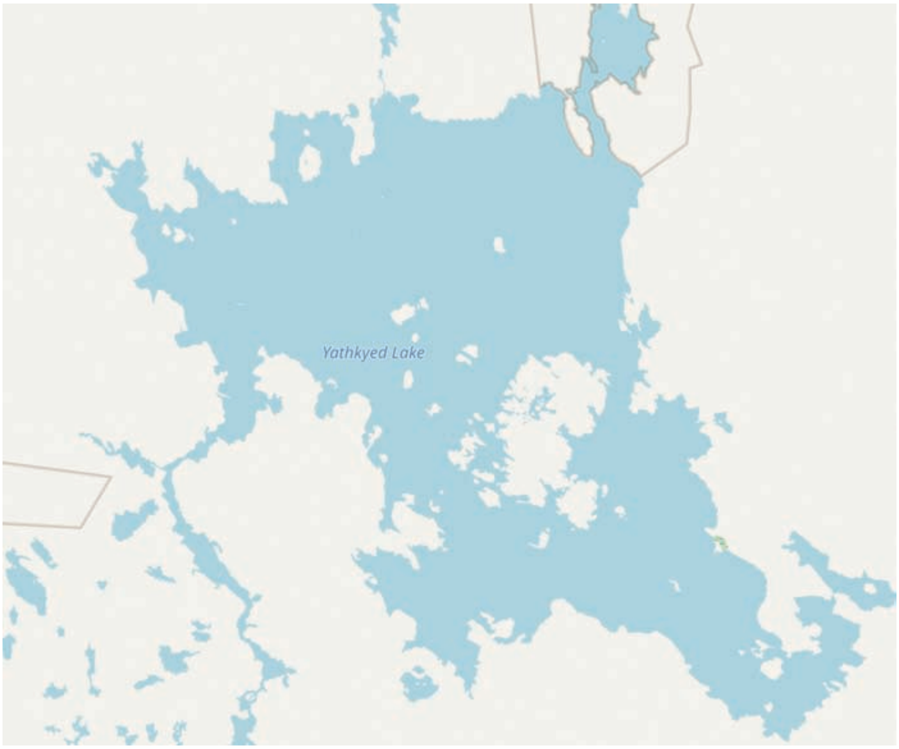


FIGURE 5.3 Encoding a lake on an island that is inside a lake that is on an island that is inside a lake. A highly nested topological relationship found in Yathkyed Lake in Nunavut, Canada

Once established, the georelational principle allows for querying the database either by location (usually in the form of an interactive map user interface) or by field values as we would commonly do in a database or a spreadsheet. The simultaneous exploration of geospatial data using either the map or the table interface is extremely powerful. But before we get into the (exploratory) data analysis possibilities afforded by GIS, we will have a look at what contributes to the popularity of GIS. In the case depicted in Figure 5.4, a table query resulted in the city of Thiruvananthapuram, India, being selected and then being marked in yellow on the map. Alternatively, we could select any of the feature geometries on the map to then display its geographic attributes in the table that is linked to the geometries by the georelational principle.

5.2 GIS MODELS

Until now, we have been very vague with respect to the geometries used to position the objects of our inquiry on the Earth's surface. We gave examples of zero-, one-, and two-dimensional features and mentioned that traditionally, the geometries are stored separately from the attributes, where we characterize the nature of the things we want to reason about. Historically, this separation made a lot of sense because we could continue to work with spreadsheets and database tables for the non-geometric components and kept the specialized geometry descriptions (as well as the coordinate system information) separate. Another advantage of this separation is that we don't have to accommodate for the many different data types as part of our table definitions.

5.2.1 Spatial Data Formats

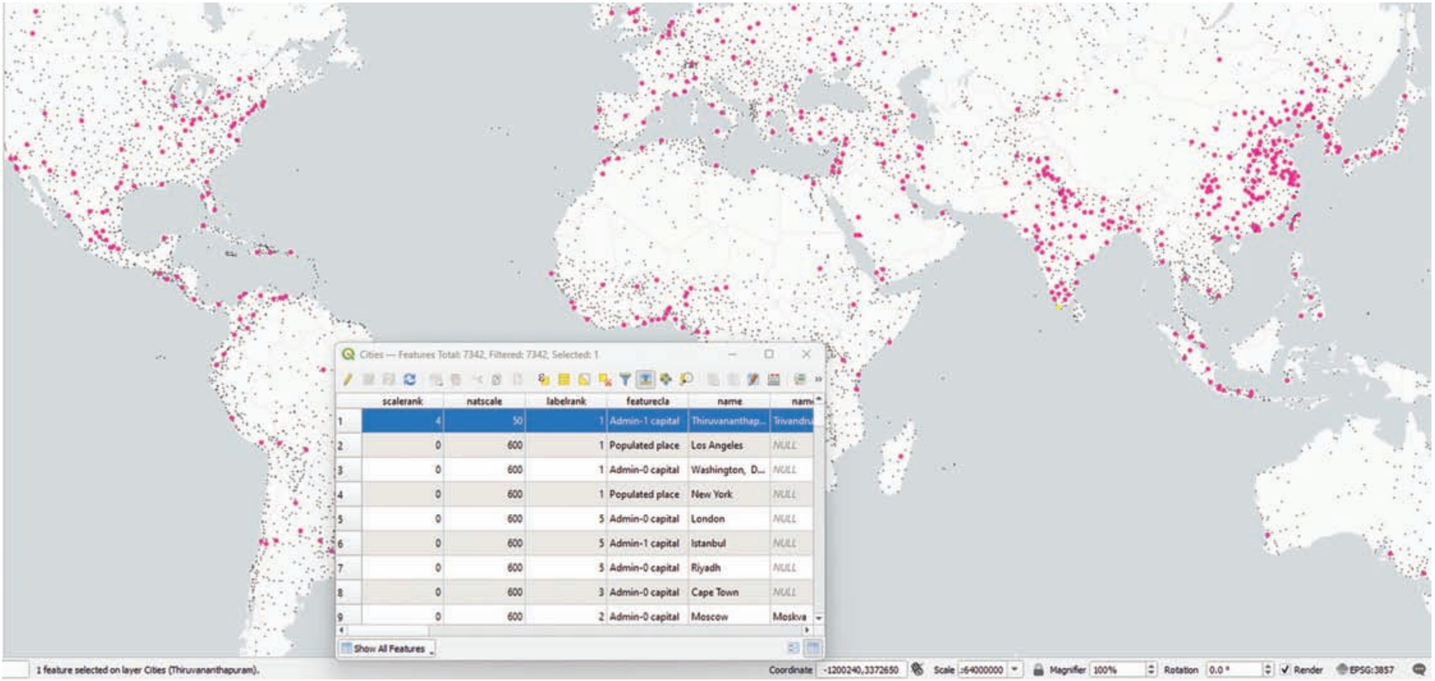
The easiest way to transition from a simple table is if the spatial reference is just a point. In that case, we may remain with a comma separated value file, where we put the point information in quotes, which allows us to store *x*, *y*, *lat*/*lon*, or even address information.

The spatial information depicted in Figure 5.4 is overly redundant, although it is common in municipal data to store the same information in multiple ways to accommodate the needs of different audiences. Things get a little more complicated, the moment our spatial reference is a linear or areal object, not to mention non-simple geometries. This is where we encounter the historic split into data formats that separate out the geometries and more modern representations that accommodate variable length fields.

The most widely used format that follows the logic of the geo-relational principle is what is misleading called the shapefile. It is misleading because a shapefile is actually a combination of at a minimum three and possibly as many as seven different files that have to be co-located in the same directory or folder and all have the same name but different file extensions. Because the shapefile is a combination of files, they are usually exchanged in the form of a .zip archive. What is confusing about the name is that one of these required components of a shapefile is a file with the extension .shp, which contains the geometry information. The other two required files are .dbf, where the attribute information is stored, and .shx, which implements the georelational principle by linking each record in the .shp file with its counterpart in the .dbf file. There are other geospatial file formats that implement the georelational principle by the same vendor (Esri) and by others but the shapefile is by far the most common one. It has been around for over 30 years and has a number of disadvantages, including but not limited to:

- Attributes are stored in a dBase file, which hails from the early 1980s and carries the limitation of the early MS-DOS operating system, namely very few data types and severely limited variable name conventions
- Geometry types are separated, i.e., points, and lines, and areas have to be stored in different shapefiles
- There is no way to store topological relationships

The shapefile format *used to be* the default in many different GIS but the above-mentioned disadvantages led to the development of a multitude of more flexible GIS data formats. Beyond the realm of GIS, markup languages are providing the basis for a number of geospatial formats that can be encoded as ASCII files (similar to the original .csv format) but now allowing to encode geometry information in the form of long strings. Both the original keyhole (.kml) as well as the geography markup language (.gml) fall into this category. Among web developers, the Javascript Object Notation (JSON) is widely popular and has spawned geospatial variants in the form of geoJSON and topoJSON. A decade ago, these would have been considered unwieldy because their plain ASCII storage causes these files to be rather



Cities — Features Total: 7342, Filtered: 7342, Selected: 1

	scalerank	natscale	labelrank	featurecla	name	nam
1	4	50	1	Admin-1 capital	Thiruvananthap...	Trivandru
2	0	600	1	Populated place	Los Angeles	NULL
3	0	600	1	Admin-0 capital	Washington, D...	NULL
4	0	600	1	Populated place	New York	NULL
5	0	600	5	Admin-0 capital	London	NULL
6	0	600	5	Admin-1 capital	Istanbul	NULL
7	0	600	5	Admin-0 capital	Riyadh	NULL
8	0	600	3	Admin-0 capital	Cape Town	NULL
9	0	600	2	Admin-0 capital	Moscow	Moskva

Show All Features

1 feature selected on layer Cities (Thiruvananthapuram).

Coordinate -1200240,3372650 Scale 1:64000000 Magnifier 100% Rotation 0.0 ° ✓ Render EPSG:3857

FIGURE 5.4 Showing a query by location vs. a query by attribute

voluminous. However, in the age of "Big Data", this does not seem to be an issue anymore, and the easy readability and their similarity to data formats beyond GIS makes them now very popular data exchange formats for geospatial data. Most open data repositories now offer geoJSON and GML as a download option.

An additional advantage of these formats is that they can be loaded into a simple text editor and parsed by non-expert GIS users. For local storage and efficient analysis, however, housing researchers should adopt a spatially enabled database. Larger organizations will probably already have their in-house DBMS, which can be spatially enabled (for free, if the DB is open-source). Smaller organizations or individual researchers are better served with personal databases that implement a DBMS in a single file. In the 1990s, this was exemplified by MS-Access but now we have specialized (and standardized) geospatial databases like SpatiaLite and building on that the GeoPackage format, see Figure 5.5. SpatiaLite is an extension of SQLite, an open source database that is built into every mobile phone, many operating systems, and appliances. The GeoPackage "is an open, standards-based, platform-independent, portable, self-describing, compact format for transferring geospatial information" (OGC, 2022).

It is now the default storage format for the widely used free and open-source software Quantum GIS (*aka* QGIS) and is suitable for all but the largest GIS implementations (for which a full-fledged DBMS is a must). Because .gpkg files implement a whole database in a single file, they are easy to share or archive. GeoPackages implement a multitude of common geometries including

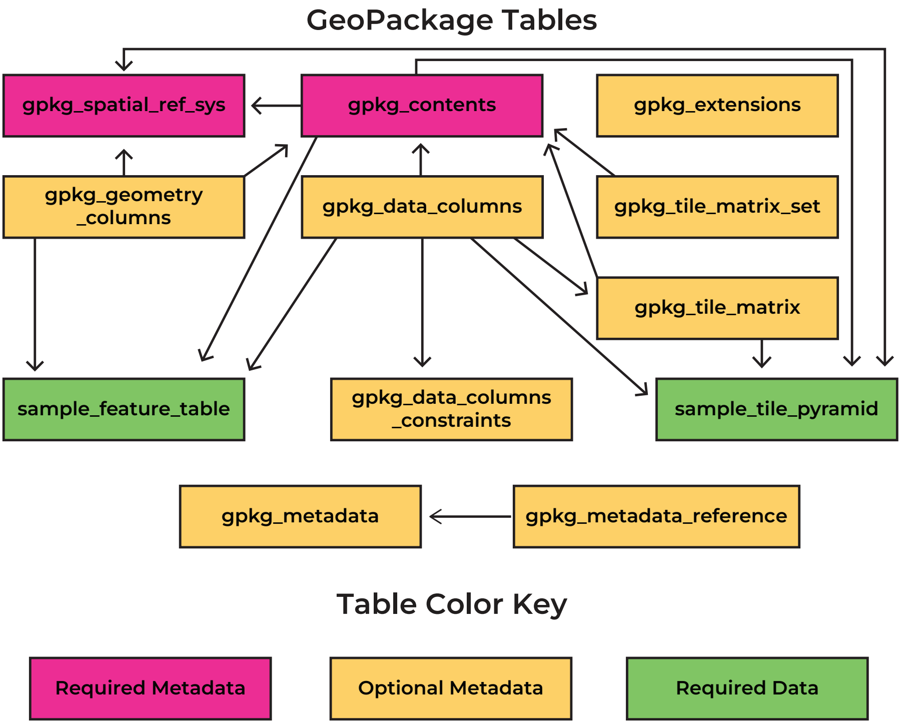


FIGURE 5.5 Required and optional components (tables) of a GeoPackage

a. Vector feature data
b. Imagery tile matrix data
c. Raster map tile matric sets
d. non-spatial tabular data, and
e. metadata that describes other stored data

Items (a) through (c) will be discussed in the following sub-sections.

5.2.2 Spatial Data Models

The first three bullet points in the content list of a GeoPackage required some elucidation, as they describe formalizations of descriptions of space that are common to *all* GIS and as such reach beyond the scope of GeoPackages alone. So far, we have always referred to points, lines, or areas as the spatial footprint of our objects of interest. In the world of Housing GIS, these types of geometries are by far the most common; in the world of GIS more generally, they are referred to as vector features. The term "feature" is used whenever we are dealing with something that has a well-defined boundary and in addition to the geometric description of those boundaries implements the geo-relational principles by adding non-spatial attributes. The term "vector" derives from the mathematical origins of encoding the *boundaries* of features. Contrary to popular conceptualizations, an area (or in GIS terms, a polygon) is not described by what is inside the area but by its boundary. The boundary of an area is made up of a minimum of three (but potentially thousands of) lines. Everything inside the boundary is taken to be uniform; there is no further differentiation of such an area, as this would require another boundary – as in the island-in-a-lake example depicted in Figure 5.3. The lines that make up the area boundary are again defined by their respective boundaries: the start and end points of each straight line. Zero-dimensional points also have a boundary themselves. So, everything in the world of vector features boils down to a collection of points, which are defined by their position relative to the origin of the coordinate system. The imaginary line from that origin to the position of a point is called a vector – hence the name vector feature and by extension Vector GIS.

Complementary to the way of conceptualizing entities in space by their respective boundaries is the notion of a field. Fields are well-known in the physical sciences: electric, magnetic, gravitational, etc. fields. What characterizes fields is their lack of boundaries. They represent phenomena that are clearly discernible but hard to fix in space. Most aspects of nature fall into this category: where is the beginning or end of a mountain, a (natural) forest, a coastline (don't forget the tides)?[1] As there are no boundaries, traditional vector geometries would be useless for describing such phenomena. GIScientists solved this conundrum by describing space, known as a raster, rather than objects in space. The term is of German origin and would typically be translated as grid. A raster/grid divides a study area into uniformly shaped and sized areas: triangles, squares, or hexagons, with square being by far the most common tessellation – although hexagons are becoming more popular as of late.

Both the remote sensing community and the GIS community have been inventing this data model in parallel. There are lots of similarities between the images in remote sensing (which, contrary to images taken with a camera or by a desktop scanner, are also georeferenced) and the grids used in raster GIS. The rationale to use rasters could either be the application (where there are few or no discernable boundaries) or the data capture instrument (a camera or similarly working instrument). Where vector features have a scale (the smallest element represented), raster datasets have a resolution (the size of each grid cell in units on the ellipsoidal surface representing Earth). For a given extent, the finer the resolution, the more cell values have to be recorded; this causes raster datasets to be significantly larger than vector files (recall that in the vector world, we record nothing about the inside of areas, whereas in the raster world, each cell has to be recorded/stored). To avoid having to work with very large files, raster, and image files (also referred to by their mathematical name "matrices") are indexed by tile pyramids (see Figure 5.6). A tile pyramid is a table that links to multiple resolutions of the same original raster layer. This is another reason to use databases because in addition to multiple vector files, a database can also store each raster and each of the multiple resolutions of a raster as separate tables that can be linked by yet another table.

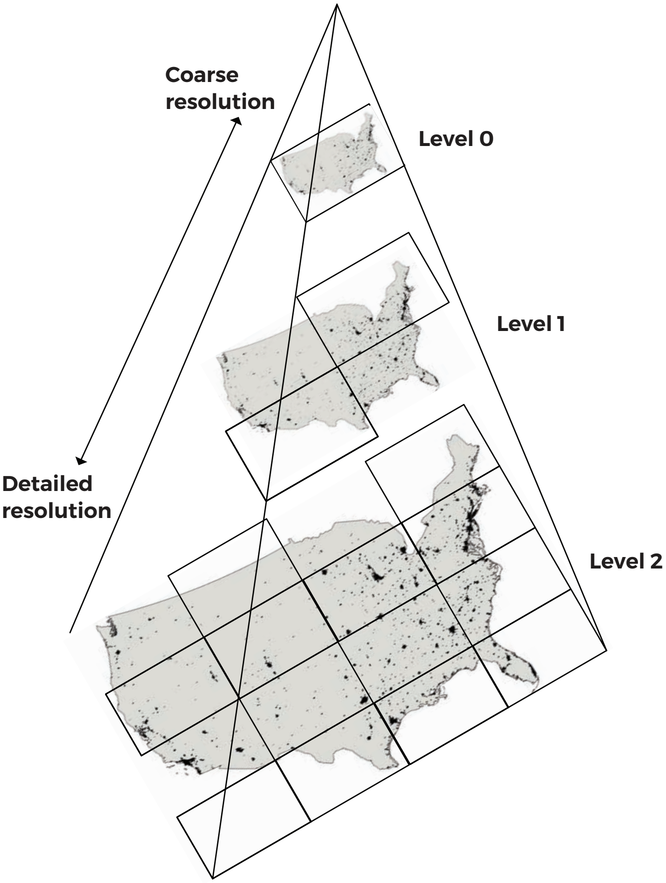


FIGURE 5.6 Unveiling the pyramid: diagram of raster tile organization with a hierarchical structure

5.3 BASIC GIS ANALYSIS OPERATIONS

Until now, we have seen two main reasons to use GIS: (i) to compile a range of datasets in a database, where we use the spatial component to index and link them by location, and (ii) to use the map interface to explore spatial relationships visually, that is to use visual cues as prompts for subsequent analysis. Many novice GIS users combine datasets in GIS to overlay them visually and then use their perceptual prowess to determine relationships between features in different layers. Combining different pieces of information on a map is a good first step – but it needs to be followed by a second, where we use the power of GIS to analytically support (or reject) our observations.

What qualifies an operation to be analytical? GIScientists have an interesting perspective on this. They distinguish between (simple) queries that retrieve an existing item from a database and analytical operations, which create something new (that did not exist in the database). The boundary between the two is fuzzy but we are on the safe side if we just check whether the result of an operation is a new geospatial dataset. If yes, then this operation falls into the analysis category. As this section is about basic GIS analysis operations only, we can now separate them into two sets of operations: horizontal and vertical. Horizontal operations usually involve only one layer and our perspective is outward bound from our object(s) of interest; they are also referred to as neighborhood operations. Vertical operations look across multiple layers and seek to determine which objects or characteristics spatially coincide; they are also referred to as overlay operations. See Figure 5.7 as a visual representation of the difference between neighborhood and overlay operations.

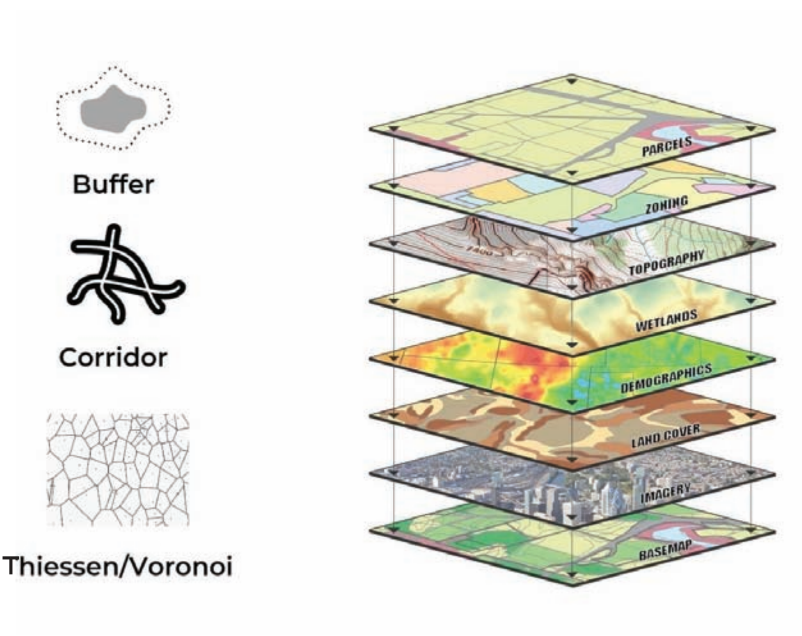


FIGURE 5.7 Neighborhood vs. overlay: diagram comparing buffer, corridor, and Thiessen/Voronoi operations illustrating spatial relationships and analysis techniques

5.3.1 Neighborhood Operations

Neighborhood operations take a location of interest (in the case of a raster representation) or the spatial footprint of a feature of interest and then have the user define an extent (the neighborhood) around it. That definition may be a simple number (units distance from the location of interest) or a heuristic where the definition of a neighborhood is a function of some attribute. Neighborhood operations then either just define a new set of features that delineate the boundaries of the neighborhood (or raster cells contained in each neighborhood) or they perform calculations on the features/raster cells within a neighborhood. This is then repeated for all features in a layer or all cells in a raster dataset respectively.

By far the most common neighborhood operation is the buffer operation, which results in a new dataset that contains all the buffers around the input features or cells. In its simplest form, the user specifies a distance, say 1,000 feet, and the GIS will create a new layer with areas of a 1,000 feet radius around the input features (e.g., bus stops). A prominent use of buffer zones in many cities is the legal requirement to identify zones around schools, where liquor (or cigarettes) may not be sold.

Another common neighborhood operation is the generation of so-called Thiessen polygons or Voronoi diagrams.[2] The input to this operation always consists of points (schools, fire stations, hospitals, etc.). Now imagine, we are simultaneously buffering all the point features (raster cells) by ever increasing distances until the boundaries of our buffer regions meet. Where they meet, we stop, but where there is still a gap, we continue our ever increasing buffer distances, see Figure 5.8 (NYC Open Data, n.d.). The process stops when there is no space left and the study area has been completely tessellated. Each input point is now surrounded by polygons that define the point's catchment area, where every location inside the catchment area is closer to the original point than to any other point. Such delineation of catchment areas is of obvious interest to every spatially aware social scientist.

So far, we assumed that as we define the distance to the input location, the new boundary is measured "as-the-crow-flies", i.e., without incorporating any obstacles. This is acceptable for phenomena that spread continuously such as noise or an air pollutant, but it is unsatisfactory for measures of accessibility. Social scientists may be more interested in taking a particular distance measure (which could also be scaled by time or safety) and then applying it along a network representing streets, or sidewalks, or transit lines.

5.3.2 Overlay Operations

Useful as they are, neighborhood operations are by far outweighed by overlay operations. As a matter of fact, for many, the whole purpose of GIS is to perform spatial overlays. This is problematic because, although the set of all different overlay operations is definitely very important and arguably makes up over 50% of all analytical GIS operations in practice, there is a world of difference between the visual overlay we discussed at the beginning of this Section 5.3 and the analytical overlay here. Let's keep in mind that analytical GIS operations *always result in new data*, not just a new map but also a new dataset that can be queried and quantitatively analyzed. Visual overlays, i.e., just displaying multiple layers in a map, is a good way to trigger research questions – but not to answer them. Rigorous housing GIS research requires

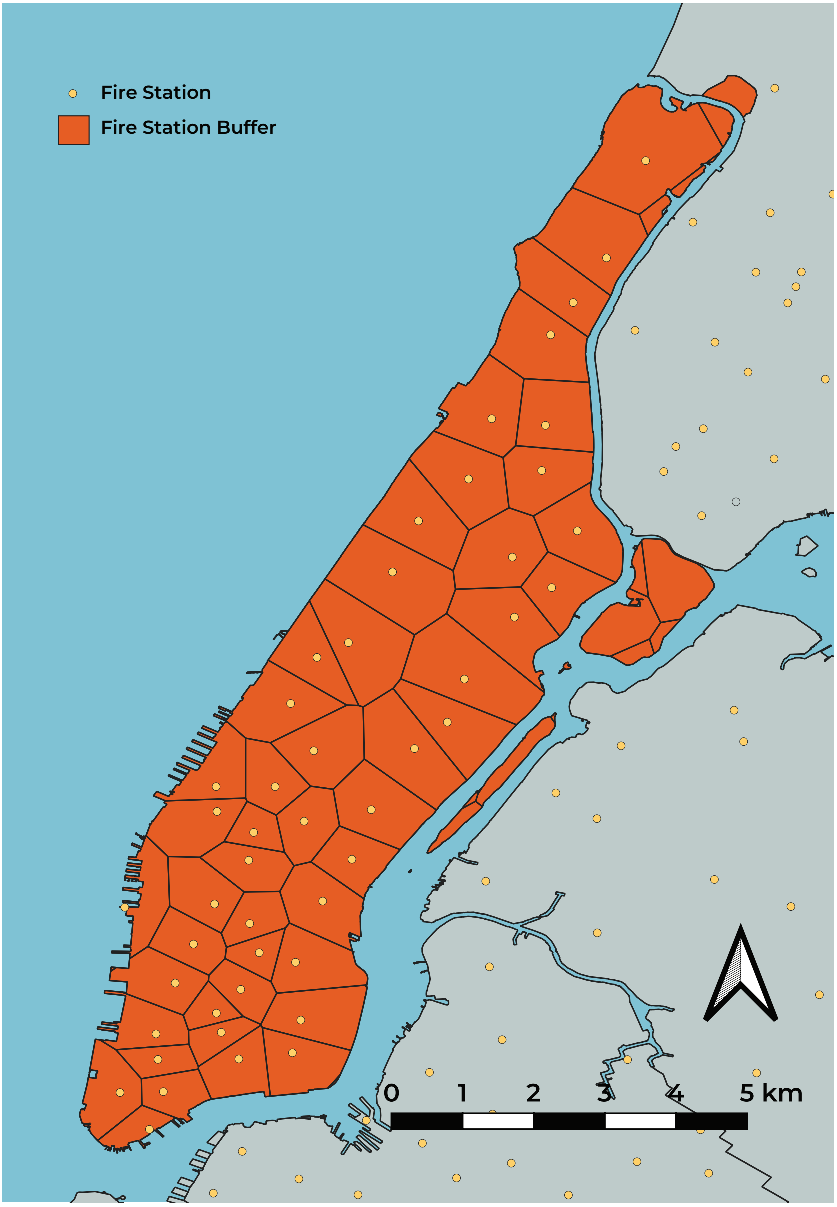


FIGURE 5.8 Fire station influence: map of Manhattan with Thiessen polygons suggesting coverage areas of each fire station

us to perform analytical overlays and understand what happens under the hood when we instruct the GIS to run one form of overlay analysis or the other.

This is perhaps best illustrated by thinking back to the MAUP (Section 4.7.1 in Chapter 4). If we want to learn about the median age of housing stock in a ZIP code area, then we have to reason about datasets that have different spatial footprints. One may be a high-resolution property database, while the other is a definition of neighborhoods (ZIP code areas may be replaced with Zillow real estate neighborhoods, or Midwestern aldermanic districts). In any case, we may now use the larger area units as a cookie cutter to aggregate the building age values. Next, we want to combine this with the fire department's incident or inspection data. The MAUP occurs whenever boundaries of one analysis unit do not coincide with the boundaries of another analysis unit. We are now overlaying the ZIP codes area data with the fire district data. In an ideal case, we can always go back to the fine-grained property data and link the building age to the inspection or incidence records at the property address level. In that case, we do not have a MAUP issue. The classic GIS overlay situation, however, is when we are trying to combine the ZIP code level data with the fire district data. In this case, we are looking at combining the two different geometries with the two different attribute datasets. Analytical overlay operations may involve point, line, area, as well as raster datasets. And as we combine different geometries, we have to look at their topological relationships to perform the analyses. What is happening under the hood is a sequence of steps that create new geometries and then subsequently new attribute records to match those new geometries. So, let's go through these step-by-step.

Overlay operations were described as working vertically, i.e., for each location, we ask what is happening here (in this layer) and what is happening at the same location in other layers. We are comparing spatially coinciding values with each other. This in turn means that for each location, we need to look up whether we are in one feature or the other (this is a lot easier in the raster world, where we do not have feature boundaries and hence can easily compare coinciding raster cells/locations). We mentioned earlier that in vector GIS, we don't say anything about the interior of polygons – they are defined by their boundaries. This in turn means that in an overlay operation, we need to determine whether we are outside, on the boundary or inside a particular feature. We then determine the same for the features in the other layer(s) and then create new features that inherit the characteristics from their respective parent features in the input datasets. The first step is to see where we are with respect to each and every feature in our input datasets.

We then determine the same for the features in the other layer(s) and then create new features that inherit the characteristics from their respective parent features in the input datasets. The first step is to see where we are with respect to each and every feature in our input datasets. We compare outsides, on-the-boundaries, and insides of all elements and thereby determine which ones are coincident at what location. The three qualitative options are defined by the topological relationships of the participating features (it does not matter how far inside or outside we are). Several researchers working with Max Egenhofer at the University of Maine and Eliseo Clementini at the University of Aquila, Italy, developed the mathematical proofs to exhaustively formalize all possible topological relationships between the boundaries of point, line, and area objects in the early 1990s.

For each of the seven groups depicted in Figure 5.9, there is a different GIS overlay operation. Each of these operations is complex; the software needs to determine what type of geometries are involved and then perform complicated geometry calculations for each and every feature of the respective layers. This is computing-intensive and can still take hours on large datasets. The results are new, and in most cases smaller geometries than in the input layers. Multiple consecutive overlay operations result in so many small geometries that they have to be followed up by some form of reclassification that is based on the most useful combination of attributes for the research question at hand.

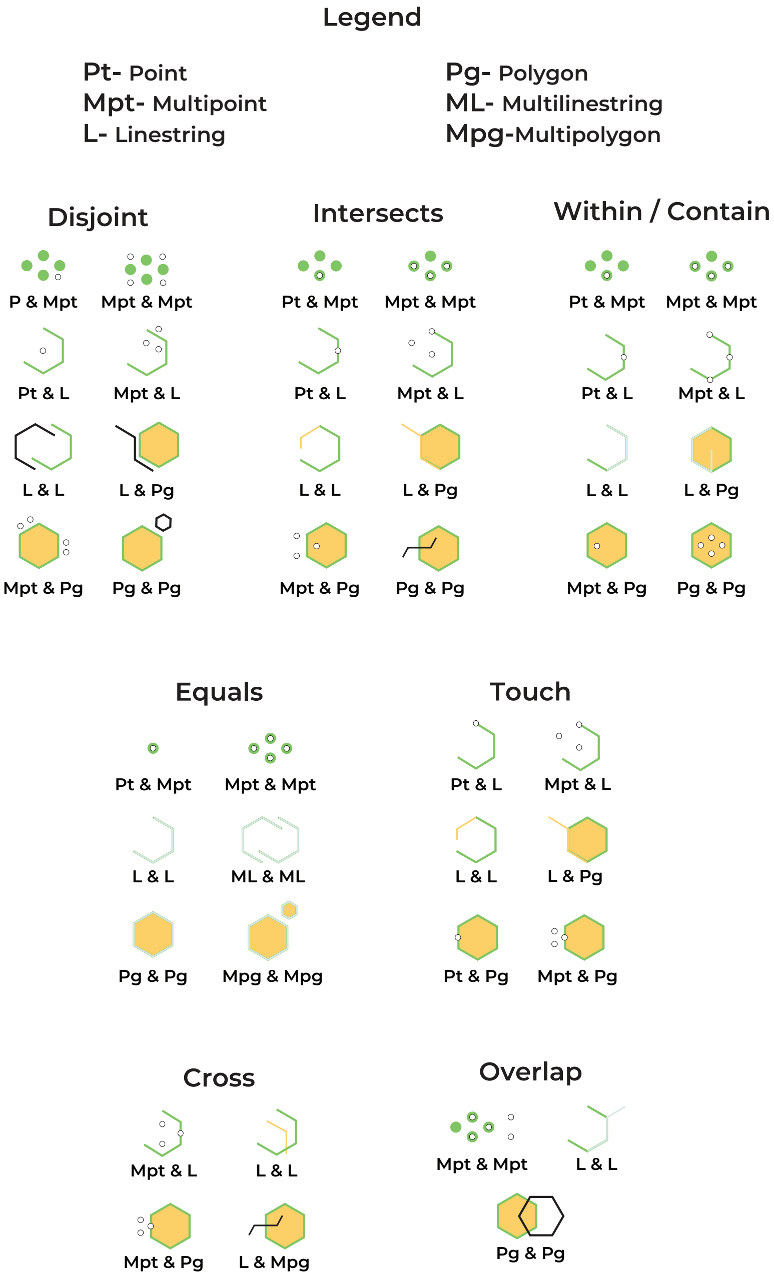


FIGURE 5.9 Exhaustive enumeration of topological relationships between 0-, 1-, and 2-dimensional geometries

Which brings us to the other side of the georelational principle? Each overlay operation involves not only the geometries but combines attributes as well. The effort we put into data cleaning and conceptual model development in Chapter 4 now really pays off because the more succinct the inputs to the overlay operations are, the easier it is to now instruct the GIS how to combine attribute values: should they be added, averaged, or reapportioned as a function of size of the areas? Overlay operations are indeed very powerful and may even be the essence of GIS.[3] With this power comes the responsibility of the housing researcher to understand the difference between the seven types of overlay operations and the need to develop a conceptual model that guides us in the choice of which operation to apply.

Between the neighborhood and overlay operations, we covered around 70% of analytical GIS operations, housing researchers are going to apply on a regular basis. Before we deal with the remaining 30%, let us have a look at how the basic analytical operations are used by housing researchers in a set of typical examples.

5.3.3 From Simple GIS Operations to Workflows

In Section 4.5 we discussed conceptual models as the foundation for a database schema. This is a good and necessary step, for if we don't have our data in place and properly organized, then there is nothing that we can apply our GIS operations to. But housing research is typically more complicated than just applying one GIS operation or the other.

At a high level, a typical GIS workflow would consist of these nine steps:

1. Define research objectives: Clearly outline the goals of the housing policy research, such as identifying areas with a high concentration of affordable housing or analyzing the impact of zoning regulations on housing development.
2. Collect data: Gather relevant data from various sources, such as census data, housing market data, zoning regulations, and land use data. This data will be used to create GIS layers and perform spatial analysis.
3. Data preparation: Clean and preprocess the collected data to ensure its accuracy and consistency. This may involve geocoding addresses, converting data formats, and standardizing attribute information.
4. Create GIS layers: Import the cleaned data into a GIS software and create layers representing different aspects of the housing policy research, such as housing prices, zoning regulations, and population density.
5. Perform spatial analysis: Use GIS tools and techniques to analyze the relationships between different layers and identify patterns or trends. For example, you might use spatial overlay analysis to determine the areas with the highest concentration of affordable housing or buffer analysis to identify the impact of zoning regulations on housing development.
6. Visualize results: Create maps and other visualizations to effectively communicate the results of the spatial analysis. This may include thematic maps, heat maps, or 3D visualizations.
7. Interpret findings: Analyze the results of the spatial analysis and draw conclusions about the housing policy research objectives. This may involve identifying areas in need of affordable housing development or recommending changes to zoning regulations to promote housing diversity.

8. Communicate results: Share the findings of the housing policy research with stakeholders, such as policymakers, housing developers, and community members. This may involve creating reports, presentations, or interactive web maps to effectively communicate the results and support data-driven decision-making.
9. Monitor and evaluate: Continuously monitor the housing market and policy changes to evaluate the effectiveness of the research and make adjustments as needed. This may involve updating the GIS layers, conducting additional spatial analysis, or refining the research objectives.

If our research question is to analyze the effect of changing zoning rules to allow for accessory dwelling units (ADUs), then step 5 above can be further broken into this sequence of GIS operations:

1. Identify zoning layers: Start by identifying the zoning layers in your housing database that are relevant to the research question. This may include layers representing current zoning regulations, land use, and existing housing stock.
2. Create a new zoning scenario layer: Make a copy of the current zoning layer and modify it to reflect the proposed changes, such as allowing ADUs in specific zones or relaxing density restrictions.
3. Overlay analysis: Perform an overlay analysis to identify parcels that would be affected by the zoning changes. This involves overlaying the new zoning scenario layer on top of the existing land use and housing stock layers to identify parcels where ADUs would now be allowed.
4. Calculate potential ADU capacity: For each affected parcel, calculate the potential number of ADUs that could be added based on the new zoning rules. This may involve considering factors such as lot size, setbacks, and maximum building height.
5. Summarize potential ADU capacity by zone: Aggregate the potential ADU capacity calculated in the previous step by zoning category or neighborhood to get a better understanding of the overall impact of the zoning changes on ADU development.
6. Analyze the impact on housing affordability: Assess the potential impact of the increased ADU capacity on housing affordability in the affected areas. This may involve comparing the potential ADU capacity to current housing demand, analyzing the potential impact on housing prices, or estimating the number of affordable units that could be created through ADU development.
7. Assess the impact on infrastructure and services: Analyze the potential impact of the increased ADU capacity on local infrastructure and services, such as transportation, schools, and utilities. This may involve using GIS tools to estimate the additional demand for these services and identifying areas where upgrades or expansions may be needed.
8. Visualize the results: Create maps and other visualizations to effectively communicate the results of the analysis. This may include thematic maps showing the potential ADU capacity by zone or neighborhood, with heat maps illustrating the impact on housing affordability, or 3D visualizations depicting the potential changes to the built environment.

Both of the above lists are fairly generic. The first one applies to virtually all GIS projects, regardless of whether they are in ecology, crime analysis, or housing research. The second list is more specific to our application area but still generic enough to be replicated, say for each neighborhood in a city – with slightly varying parameters as our requirements change from one location to another. It is worthwhile mentioning that the operations themselves are very basic; their impact derives from the repeated application of the same small set of basic operations to intermediate outcomes. If we can save the sequence of processing steps as a model that can be executed with a single click, then we (i) avoid the tedium of repeated the same steps again and again, (ii) ascertain that when we run the model again it can be compared with previous model runs because the steps are guaranteed to be the same, and (iii) we can share this model with a colleague. In information programming terms, this would be called creating a function. In the world of GIS, this model creation is referred to as geoprocessing (a term coined by the company Esri) or just plain processing (in the world of free and open source GIS).

The reason we began this section with a nod toward our discussion of conceptual models in Section 4.5 is that we should treat the development of such processing workflows as the other side of the same conceptual modeling coin. One of the authors of this volume has built his career on the development of tools for such workflow modeling. Simple models can be built with GIS-internal tools but complex models that link to larger institutional (and non-spatial) workflows would benefit from using either the Unified Modeling Language (UML) or the software that implements the standards of the Business Process Modeling Notation (BPMN). However, regardless of whether we sketch out our workflow on the back of a napkin (not a bad idea!) or using a formalized language, the development of a workflow sequence (i) helps to clarify in one's mind what exactly it is we are trying to accomplish with our GIS work, (ii) helps us to document our workflows both for the sake of communicating it in a final report as well as to build institutional knowledge, and (iii) develop a small library of standardized workflow models that are unique to the enterprise we are working for and can be deployed with the push of a button to anyone with a barebone knowledge of GIS.

5.3.4 Basic GIS Functionality in Housing Policy Research

Section 5.5 will provide some in-depth examples for GIS use in housing policy research. This subsection is a prelude to provide the reader with a few practical examples of the otherwise rather abstract and technical discussion of basic GIS analysis operations. We will illustrate the application of neighborhood and overlay operations with two commonly asked questions: (i) is there a relationship between building permits and gentrification, and (ii) are rents higher or lower near transit stops? The first question can be answered with overlay operations only, while the second question requires a combination of overlay and neighborhood operations.

The first question is also a fine example for the importance of conceptual models because depending on how we conceptualize the term gentrification, we would try to capture this phenomenon with a range of different variables. Even something as plain as building permits deserves a little further scrutiny because permits for new construction typically do not cause displacement (unless it is preceded by demolition), while building alterations often require tenants to vacate at least temporarily. In an aspatial world, we would just look for the respective values of whatever variables we found to be representative of our research question and then look at trends on a city-wide or state-wide scale to determine whether there is a correlation between the number of permits and the gentrification indicators. With GIS, however, we are aiming to capture the variations in space. Where do we have how many buildings permits and where do we observe what gentrification indicator values, be they rental price increases above the regional average or the percentage of people who did not live in a neighborhood 5 years ago? Each variable becomes a GIS layer that allows us to depict the local or regional differences. Assuming that the gentrification variables are combined into a summary indicator, we can then perform an overlay between the building permits data (typically point data that we could summarize to the level of area units that we measure gentrification in) and the gentrification layer. This spatial perspective will then provide us with evidence for where there is the presumed relationship and where there is not.

The second question takes a horizontal perspective, where again, we have to consult our conceptual model to determine what "near transit stops" means. What mode of transit should be included and how far do we anticipate the influence to reach? In theory, we might even do without any preconceived notion of horizontal reach because in a perfect world, we would need to only map the spread of rental rates and if there is a relationship, then we should observe hot or cold spots (troughs and peaks in a 3-dimensional representation) wherever there is a transit stop. But chances are that the spatial relationship between our two observables varies across the study area and so we typically define catchment areas around each transit stop (either simple buffers as-the-crow-flies or along a road network depicting temporal isolines) and then compare the average rents inside the catchment areas with those outside.

In Section 5.5, we will delve a little deeper into the range of GIS analyses that are available to housing policy researchers. But before we go there, we need to discuss the role of visual communication that accompanies any GIS analysis.

5.4 GIS FOR MAPPING AND VISUALIZATION

One of the main attractions of GIS is its ability to engage the housing policy researcher through its interactive map-based user interface. It is this visual representation and the opportunity to interactively explore spatial relationships on a map that sells GIS to larger audiences. Visualization occurs at all stages of the GIS process. Whenever we receive a new dataset, we should look at it both from a descriptive statistics perspective as well as display the data on a map. In both instances, a cursory (but purposeful) look at the data will give us clues about their usefulness; but what is unique about the map is that it is prone to draw us into exploring spatial relationships. The map will provide us with situational context and prompt us to look for patterns. This is built into us humans; we may actually detect patterns that turn out not to be statistically relevant – but this is what the analytical part of GIS is for. Most people looking at a map will try to reconcile what is displayed with what they know about the place. The mere display on a map will either confirm what we know or will invite questions about whatever surprises us, see Figure 5.10 (US Census, n.d.; NYC Open Data, n.d.).

This process of visually making sense of the data should be done for each dataset individually, and then by looking at the relationships between the different datasets. Part of the mythos of GIS is that each dataset becomes its own map layer and that we can stack map layers on top of each other to then visually explore the relationships

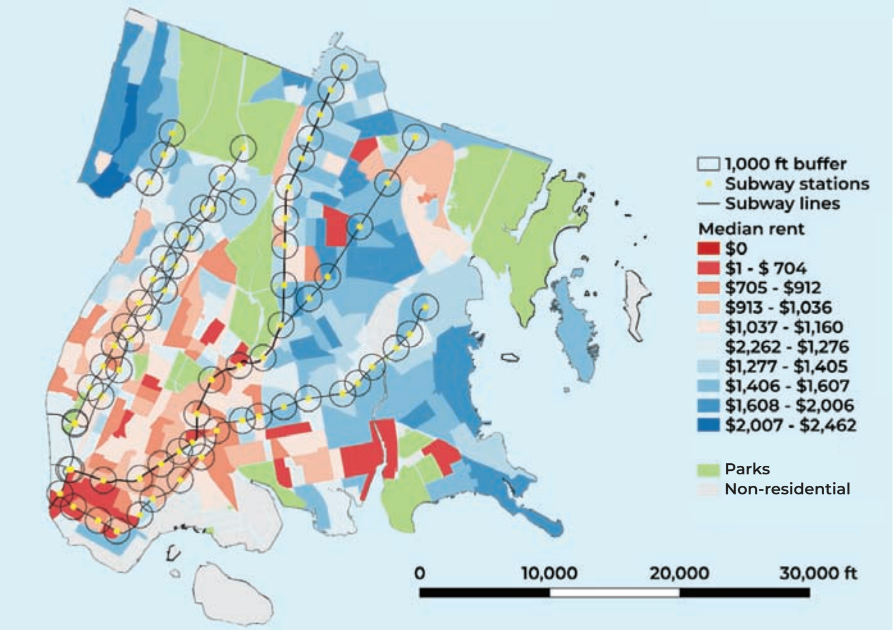
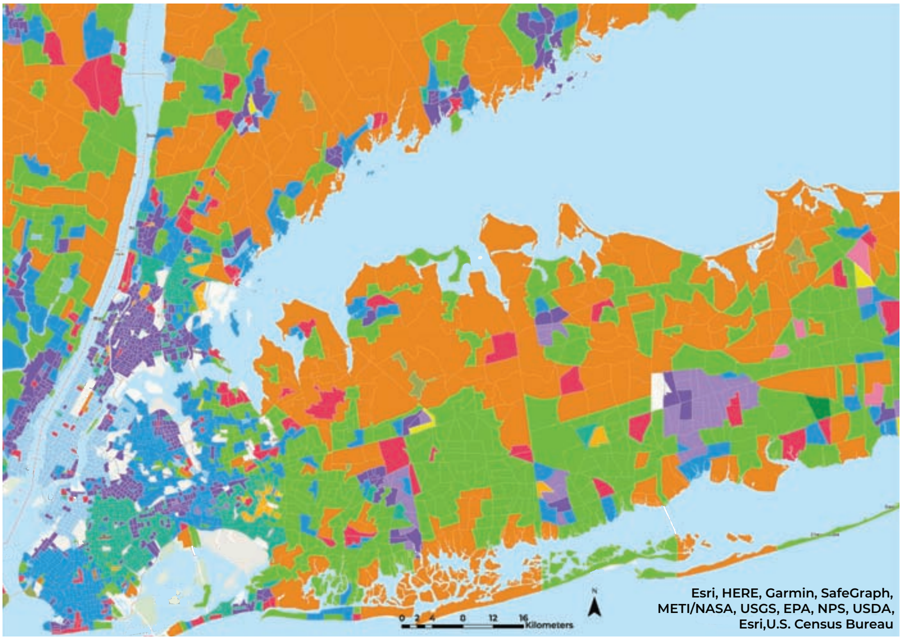


FIGURE 5.10 Rent and transit nexus: GIS map showcasing the relationship between median rent and proximity to subway stations using multiple layers

among the map elements *across* the layers. This takes us back to the notion of conceptual models discussed in Section 4.5 as well as the basic GIS examples at the end of the previous section. Which data points coincide spatially or are in close vicinity to each other? Is there a relationship between building permits and gentrification? Are rents higher or lower near transit stops? The visual exploration will again help us to generate research questions and to check our initial assumptions (which will have to be confirmed using the analytical methods of Sections 5.3 and 5.6). The ability to jump back and forth between the table and the map interface and to have these linked through the georelational principle is one of the big selling points for GIS in housing research.

5.4.1 Tapestry Data

One of the best examples for putting our own data into context and then applying spatial reasoning is Esri's tapestry segmentation data, a well-developed example of geodemographics that identifies 67 different spatialized market segments. Using data clustering and data mining techniques (partially discussed in Section 5.6), Esri delineated contiguous areas (which they call neighborhoods) throughout the United States, where the resident population falls into one of the 67 euphemized demographic categories listed in Table A.7, located in the appendix. Now, a serious housing researcher will compile contextual data herself rather than relying on the marketing-oriented tapestry segmentation data. However, for GIS students, this represents an excellent example of how to make sense of the housing geography of a place – especially if one is not a local. See Figure 5.11 for a map made with Tapestry Data.



Esri, HERE, Garmin, SafeGraph, METI/NASA, USGS, EPA, NPS, USDA, Esri,U.S. Census Bureau

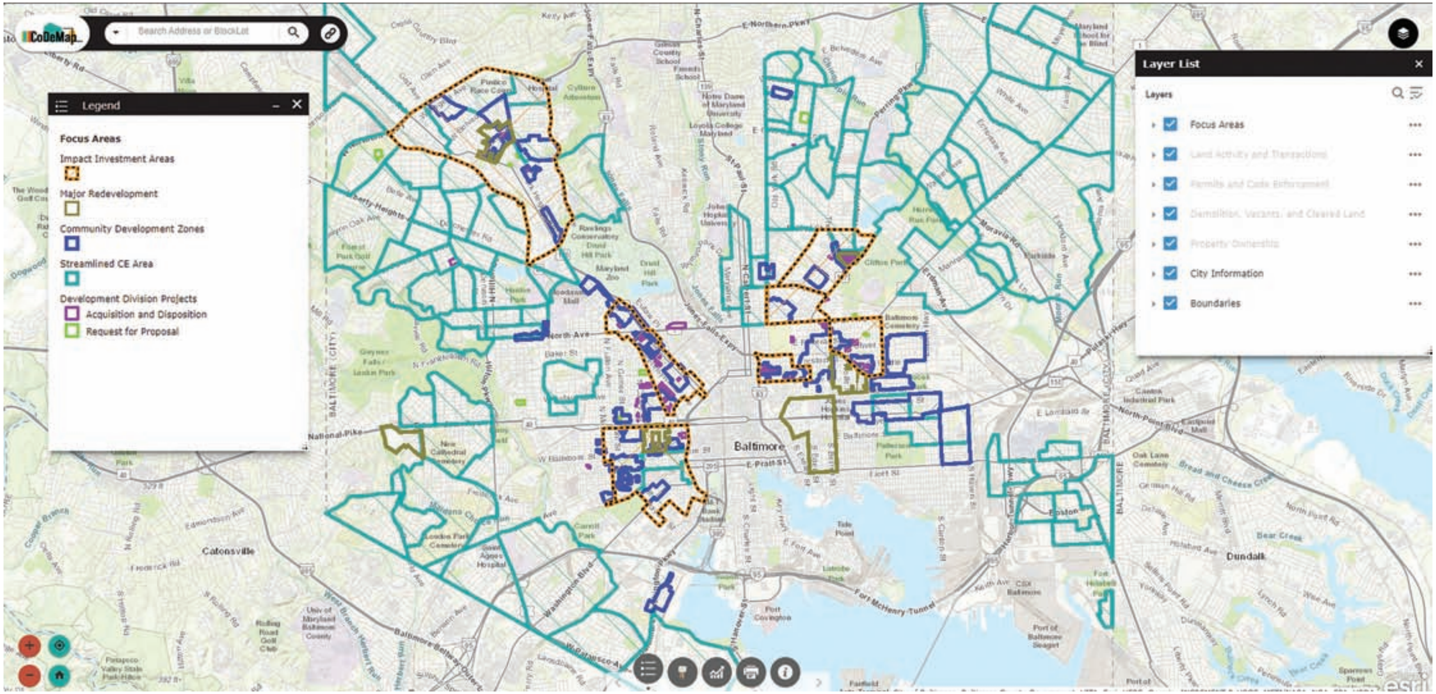
FIGURE 5.11 Market segment clusters in the Metro NY tapestry data

5.4.2 Data and Information Visualization

This volume is based on the premise that spatial differentiation matters and all the maps in this chapter so far are an illustration of the advantages of GIS when it comes to analysis. But as discussed in Section 5.2 policies need to be communicated and maps are a natural ally of the housing researcher – if deployed conscientiously.

Take Baltimore's online *Community Development Map* (CoDeMap), for instance. It visualizes housing needs in the city, neighborhood by neighborhood. CoDeMap is a central point of access for the housing department's numerous databases with everything from citation data to a property's permit history. It has evolved from a housing code enforcement tool to a platform that provides insights into housing, community development, and property datasets at the citywide, neighborhood, block, and parcel levels. It is this double function of serving both inward-facing city employees to link data across different repositories to answer specific questions, and serving the public that displays the power of GIS.

On the inward-facing side, CoDeMap can display a census block or parcel level to reveal foreclosures, open work orders, outstanding violations, property types, vacancies, ownership types, and more. Having all key data in one place also allows staff from other city departments to *see* and understand housing policies. Much of this is now shared with nonprofit organizations, neighborhood associations, and developers, who have received free training sessions that allow them to explore the riches and help to create an equal playing field when it comes to discussing new development plans. GIS visualizations (maps as well as the storyboards of the following subsection) are an immensely effective communication, discussion, and public engagement tool, Figure 5.12 (Baltimore DHCD, 2023), which highlights Baltimore's Community Development Map.



Legend

Focus Areas

Impact Investment Areas

Major Redevelopment

Community Development Zones

Streamlined CE Area

Development Division Projects

□ Acquisition and Disposition

□ Request for Proposal

Layer List

Layers

☑ Focus Areas

☑ Land Activity and Transactions

☑ Permits and Code Enforcement

☑ Demolition, Vacants, and Cleared Land

☑ Property Ownership

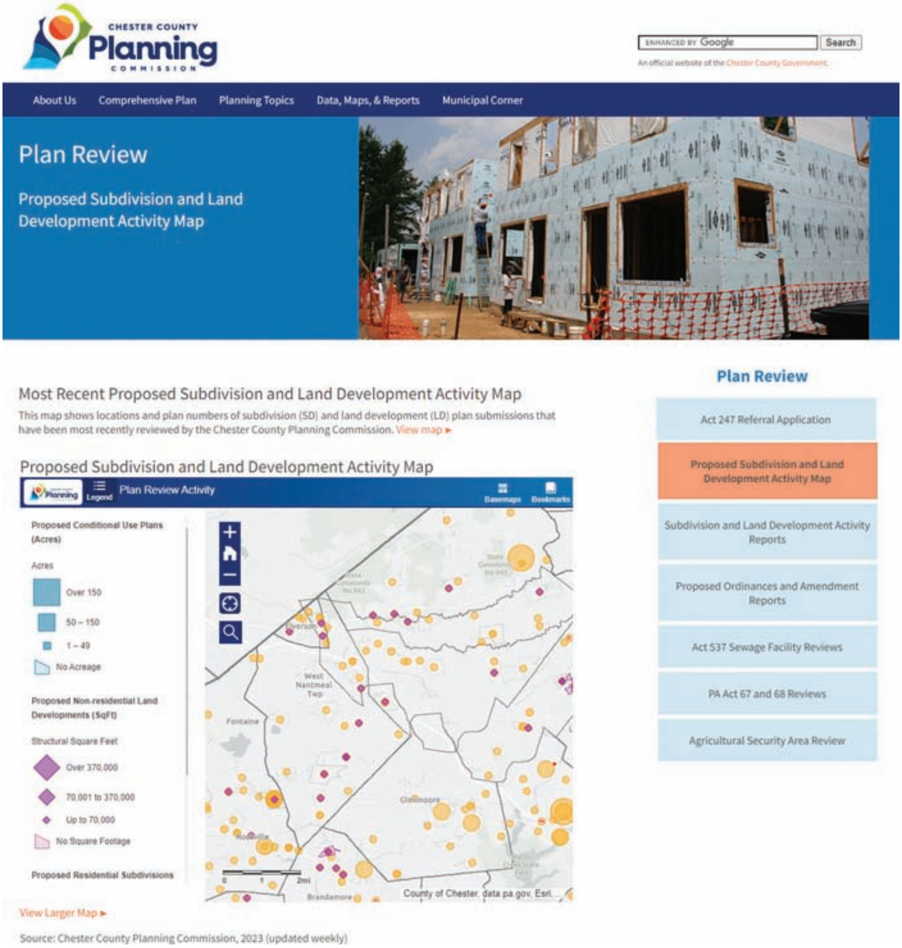
☑ City Information

☑ Boundaries

FIGURE 5.12 Community development map highlighting areas of redevelopment, development zones, impact areas, streamlined regions, and ongoing projects

5.4.3 Communication to Different Audiences

Where the previous subsection concentrated on visualizing the co-location of different aspects of a planning decision in a desktop environment, we are now discussing examples of taking interactive GIS displays to the Web. Public outreach is a legal requirement for virtually all housing policy decisions. Figure 5.13 (Chester, 2023) could hail from a traditional static local planning department webpage. But this is just the luring entry point to a website that then engages the visitor with its ability to query the system based on their own home address (Figure 5.14, (Bucks, 2023)). It is easy to engage citizens if they are given the means to find out what is happening in their vicinity. Northern Kentucky's *Link GIS website* rivals any popular social media site with its storymaps, a mashup of text, background photos, videos, and interactive maps that we introduced in Section 3.6. By translating each (GIS) project into an *engaging story*, Link-GIS keeps justifying its existence to



An official website of the Chester County Government.

Most Recent Proposed Subdivision and Land Development Activity Map

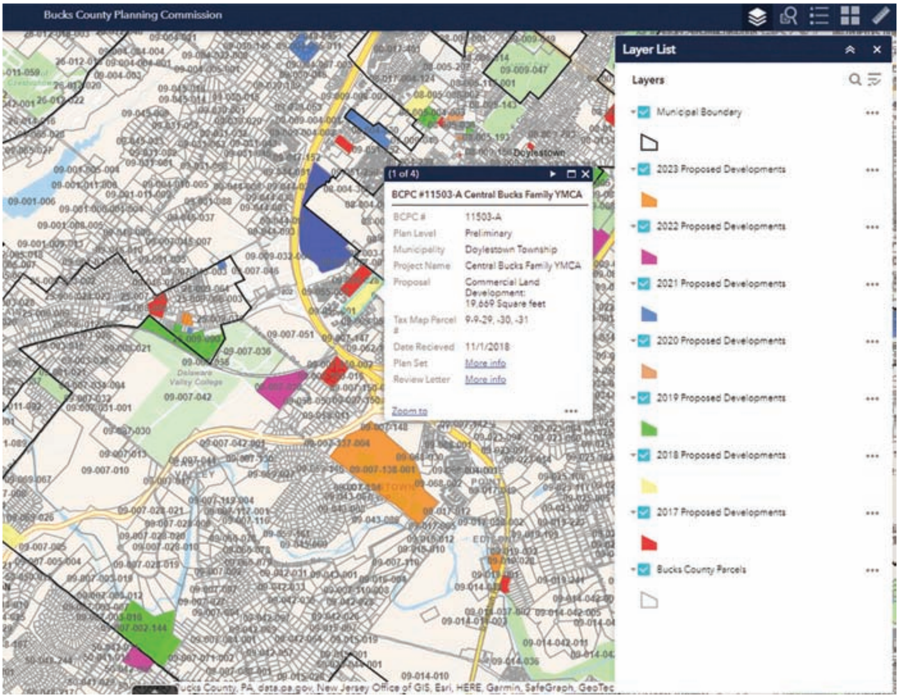
This map shows locations and plan numbers of subdivision (SD) and land development (LD) plan submissions that have been most recently reviewed by the Chester County Planning Commission. View map ▸

Proposed Subdivision and Land Development Activity Map

Source: Chester County Planning Commission, 2023 (updated weekly)

CHESTER COUNTY
Planning
COMMISSION
ENHANCED BY Google
Search
About Us
Comprehensive Plan
Planning Topics
Data, Maps, & Reports
Municipal Corner
Plan Review
Proposed Subdivision and Land Development Activity Map
Planning
Legend
Plan Review Activity
Basemaps
Bookmarks
Proposed Conditional Use Plans (Acres)
Acres
Over 150
50 – 150
1 – 49
No Acreage
Proposed Non-residential Land Developments (SqFt)
Structural Square Feet
Over 370,000
70,001 to 370,000
Up to 70,000
No Square Footage
Proposed Residential Subdivisions
West Nantmeal Twp
Fontaine
Glenmoore
Rockville
Brandamore
0
1
2mi
County of Chester, data.pa.gov, Esri
View Larger Map ▸
Plan Review
Act 247 Referral Application
Proposed Subdivision and Land Development Activity Map
Subdivision and Land Development Activity Reports
Proposed Ordinances and Amendment Reports
Act 537 Sewage Facility Reviews
PA Act 67 and 68 Reviews
Agricultural Security Area Review

FIGURE 5.13 An attractive entry point to an online GIS


Bucks County Planning Commission
Doylestown
Delaware Valley College
Layer List
Layers
Municipal Boundary
2023 Proposed Developments
2022 Proposed Developments
2021 Proposed Developments
2020 Proposed Developments
2019 Proposed Developments
2018 Proposed Developments
2017 Proposed Developments
Bucks County Parcels
(1 of 4)
More info
Zoom to
Bucks County, PA, data.pa.gov, New Jersey Office of GIS, Esri, HERE, Garmin, SafeGraph, GeoTec

BCPC #11503-A Central Bucks Family YMCA

BCPC #	11503-A
Plan Level	Preliminary
Municipality	Doylestown Township
Project Name	Central Bucks Family YMCA
Proposal	Commercial Land Development: 19,669 Square feet
Tax Map Parcel #	9-9-29, -30, -31
Date Recieved	11/1/2018
Plan Set	More info
Review Letter	More info

FIGURE 5.14 Most public visitors to an online GIS will first check out what the GIS has to say about the vicinity of their home

taxpayers while providing a public relations service for the whole community. This is basically a blog like millions of private blogs. But it keeps local citizens in the know and is an easy to link physical with online communities as in the *storymap of the Excelsior neighborhood in San Francisco*, that has been archived by Stanford's University Map Library and may hence be accessible for many years to come.

As discussed in Section 3.5, Esri's storymaps provide a convenient one-stop for creating such effective map-based means of communication. But there are free and open-source alternatives such as *MapStore*.

5.5 GIS FOR HOUSING POLICY RESEARCH

In Section 5.3, we introduced basic GIS analysis techniques in the abstract. This section will illustrate the application of these basic techniques with four examples of typical GIS use in housing policy research. Today more than ever, successful public policy depends on high-quality data and the technology that communicates its meaning effectively. Beyond the rational application of scientific or systematic methods, public policy is about values and how values affect, and are affected by, policies. This requires the delivery of credible information in a transparent, understandable form not only to decision makers responsible for adopting policy, but also to various categories of stakeholders whose behavior will be impacted in some way by the policy's implementation.

In order for public policies to be successful, it's important to have good data and technology that can clearly explain what the data means. Public policy isn't just about using science and systematic methods. It's also about values and how those values are impacted by policies. This means that people who make decisions about policies and people who are affected by policies need to have access to reliable and easy-to-understand information.

Anderson (2015) identifies five stages in the policy process:

1. Problem identification
2. Formulation
3. Adoption
4. Implementation
5. Evaluation

Our examples will deal with all of these, but special emphasis will be given to the use of GIS to determine where and when policies are needed, the formulation of public policies, the implementation, and evaluation.

5.5.1 Using Cadastral Maps for Problem Identification in Housing Policy Development

Cadastral applications were among the first uses of GIS combining the legal records (attribute data) with the surveying maps – a quintessential example of the georelational principle. Taken by itself, cadasters are little more than repositories with no need for any kind of analysis. These are hyper-local datasets that often are not public because smaller municipalities cannot afford to have their own GIS departments and are using private contractors to develop and maintain a GIS-based cadaster. Increasingly, however, say with the support of their counties, these datasets are being made public and can be used as input for interesting housing-related analyses.

Regardless of provenance, all cadastral datasets have information about the owners, see Figure 5.15 (MapPLUTO, n.d.). Just mapping the top ten landlords makes for interesting insights. Often, these are institutional (governments, churches, and universities) that have an oversized influence on land use planning decisions, but as of late these also include non-traditional landlords such as investment companies.

A second common attribute in a cadastral database is the building age, see Figure 5.16 (MapPLUTO, n.d.). Depending on whether the data has been reconciled with the buildings department (responsible for permitting), this provides valuable information about the nature of the housing stock, from insulation to lead pipes or paint or climate change resiliency.

The number of floors of a building provides useful input to both attempts at neighborhood densification as well as acting as an indicator for the potential for solar roofs (very few buildings with more than five floors have a sloped roof, suitable for the installation of photovoltaic panels; a more thorough analysis would then include aerial imagery, from which one could discern the direction in which a roof slopes, as well as whether it is shaded by trees), see Figure 5.17 (MapPLUTO, n.d.).

In cities that have used GIS for cadastral applications for a while, tax lot change analysis provides valuable insights into the effect of housing policies, see Figure 5.18 (MapPLUTO, n.d.). Information derived from a simple change analysis includes subdivisions, ADUs, zoning changes, etc. The uninitiated would think that all of this can be derived from a spreadsheet as well (basically the attribute component of GIS data) but the crucial information missed by that approach is the determination of "where".

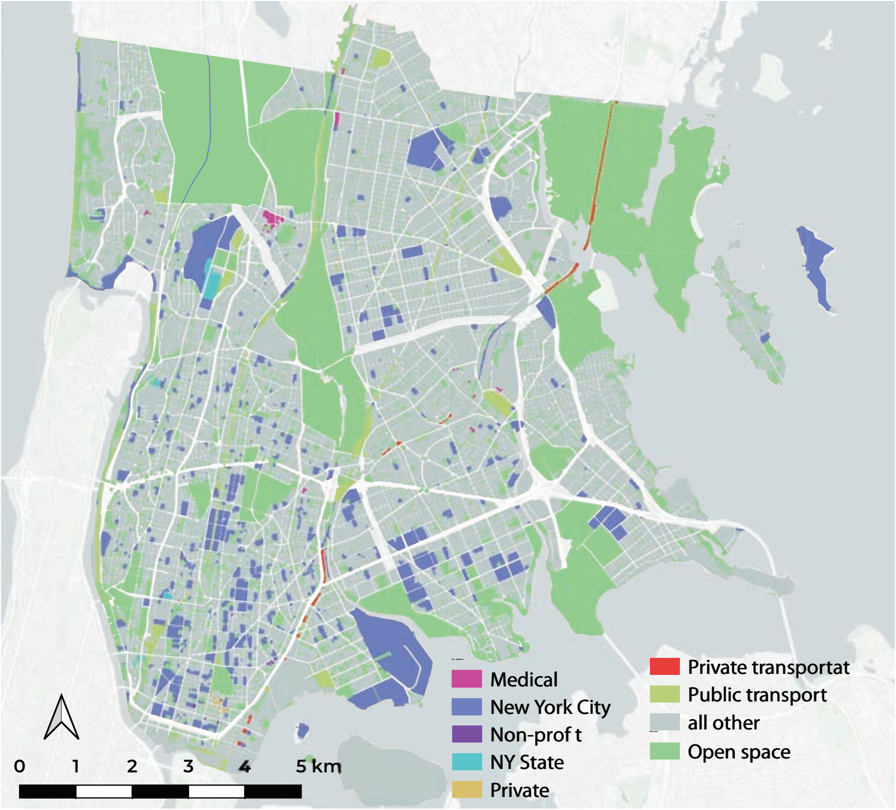


FIGURE 5.15 Bronx largest property owners

One and the same policy (change) might have very different effects in different parts of a jurisdiction.

Our last example of practical uses of a GIS-based cadaster is the socio-ecological analysis of vacant lots, see Figure 5.19 (MapPLUTO, n.d.). As many municipalities are running out of space for new housing, vacant lots offer at first sight an obvious choice for new developments. But there are always any number of reasons why a lot has not been developed. It may serve as an institutional land bank, it may be in a flood zone, it may be a brown field, or it may just be too small to warrant development without razing buildings on neighboring properties. All of these reasons could be found in a GIS database. It is the linchpin for asking questions beyond the narrow scope of the original creation of the database. This, then, is the argument for establishing such a database in a central IT department which has the capacity to link datasets across functional boundaries.

5.5.2 Using GIS to Formulate and Adopt Housing Policy Changes: Gentrification

Understanding displacement is critical given the housing crises around the country: rising rent burdens, homelessness, loss of rent-regulated housing, public housing

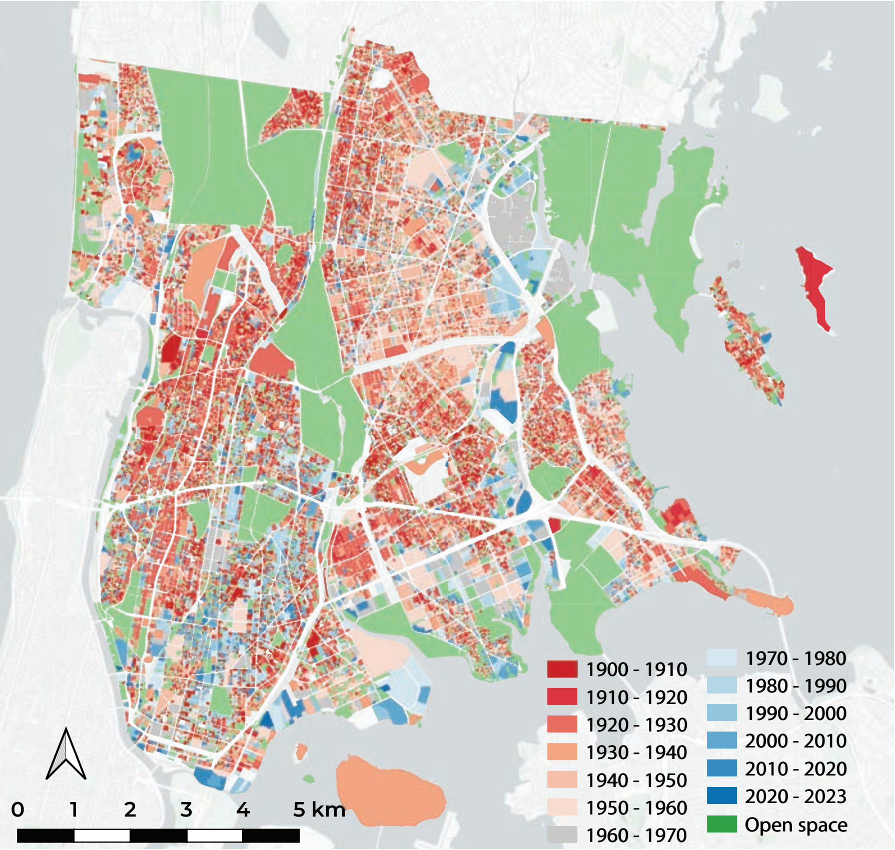


FIGURE 5.16 Bronx building age

deterioration, and more. We saw in the discussion behind the data for Figure 5.15 that non-for-profit housing initiatives are among the largest property owners in The Bronx, NY. With new federal policies like Opportunity Zones and such local actions that seek to harness market-rate development to boost the supply of affordable housing, it is time to look more carefully at displacement. A popular measure of gentrification is the increase in home values or apartment rents. The problem with that is that property values are almost always going up (everywhere). So, the question then is whether the costs have been going up in a gentrifying neighborhood more than in comparable neighborhoods nearby (with the notion of "nearby" itself being a contentious issue). Slightly more sophisticated is the question of changes in housing affordability (see Chapter 4) and again, its relationship at one location compared to another. At the heart of the gentrification debate, however, is the notion of displacement. The US Census Bureau publishes census tract-level data in response to the question "have you lived in this [area unit] 1/5/10 years ago?" If the answer leans heavily towards shorter time spans, then this may be an indicator for gentrification in a narrower sense. On the other hand, there are numerous neighborhoods around the

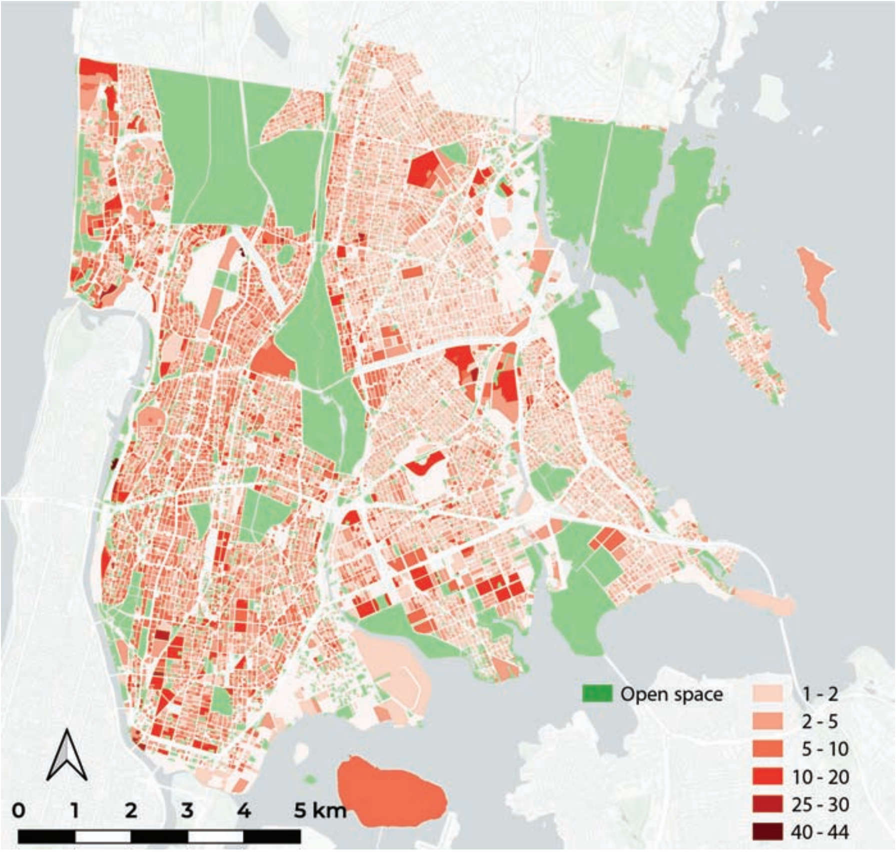


FIGURE 5.17 Bronx building heights

country (and by the way not limited to urban areas) that have always been transitory, i.e., they serve as landing points for immigrants who then move on after a few years. Even racial or ethnic changes may then be due to international causes and are not suggestive of gentrification.

The following map (Figure 5.20 (US Census, n.d.)) characterized neighborhoods as vulnerable to gentrification if housing sales prices or rent <80% of median, *and* any three of the following four can be observed:

- % low-income households > regional median
- % college educated < regional median
- % renters > regional median
- % nonwhite > regional median

We can then create categories of gentrification by comparing 2000 Census data with 2020 Census data. If a census tract had low-income communities in both years but experiences changes in any of the other bullet points, then this signals ongoing gentrification. If in addition to that, the census tract moved from the low-income to a

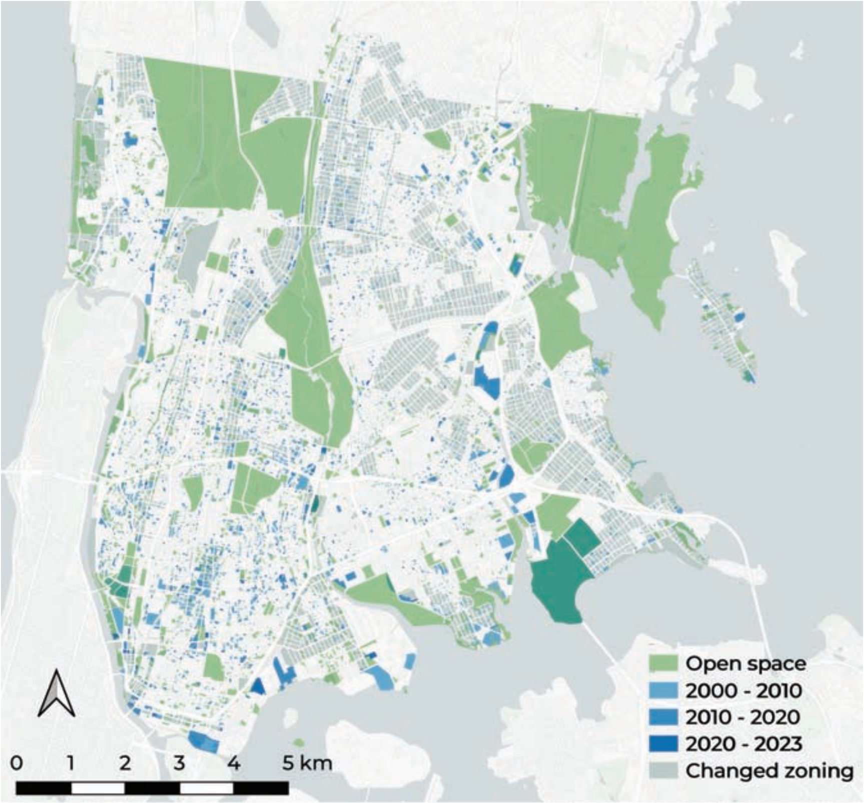


FIGURE 5.18 Bronx zoning change

middle or high-income class, then this represents an advanced stage of gentrification. “At-risk” are neighborhoods, where the only change so far is above the regional median rise of rents or median property value.

In 2019, some 20% of low-income households or 293,410 people in The Bronx live in low-income neighborhoods at risk of or already experiencing displacement and/or gentrification pressures. We represent as “missing data” those census tracts, where population counts are smaller than 500 residents or the Census Bureau’s coefficient of variation suggest a high degree of unreliability.

All of these considerations, however, will only discern the phenomenon after the fact. If gentrification is to be avoided or at least slowed down, then we need to look for indicators of potential future gentrification. A change in amenities (from new green spaces to new transit options (Checker, 2011; Chava and Renne, 2022) may serve as a harbinger of future gentrification. The cumulative effect analysis under the California Environmental Quality Act is a fine example of the utility of having not only a GIS database but, as discussed in Section 3.5.5, also a set of formalized workflows that check for interaction effects of past and present administrative actions (see Figure 5.21 (adapted from Association of Environmental Professionals, 2022)). See the section on GIS challenges in the following chapter for more on geospatial workflow management.

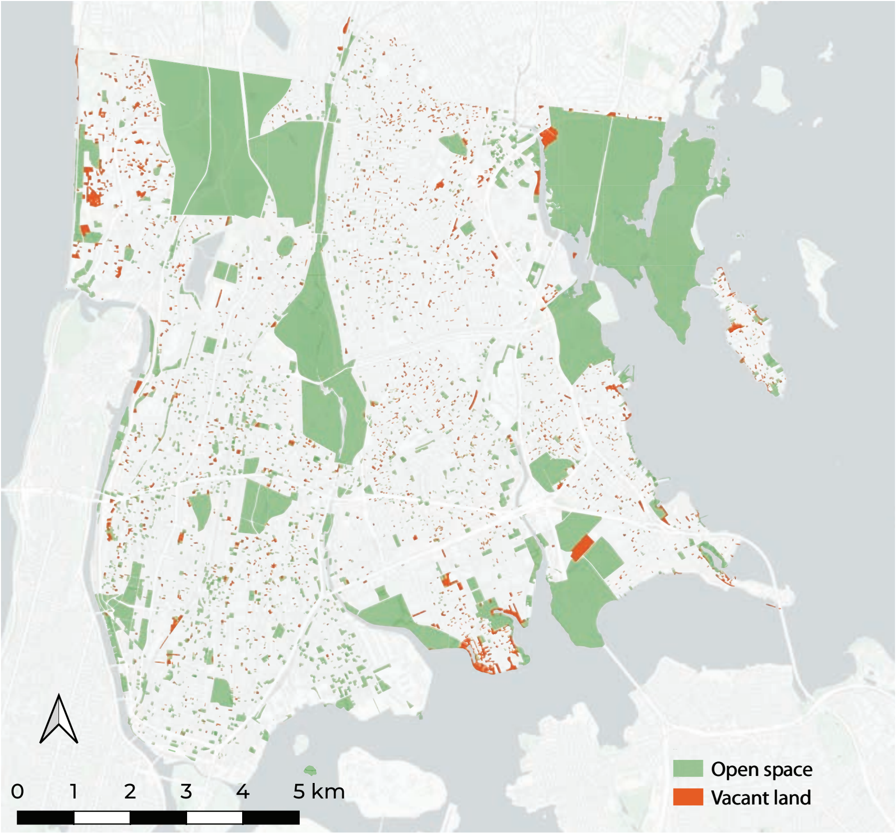


FIGURE 5.19 Bronx vacant land

5.5.3 Using GIS to Evaluate Housing Policy

The term "evaluation" can be applied in a number of different contexts. It may be interpreted as evaluating a situation to understand the severity of a problem, in other words, a needs assessment, or it may be used to evaluate a policy that was established to address the problem. We are going to discuss an example of each in the following pages.

If we are trying to understand the demand for housing in a given area, then we can, following Webster (1993) distinguish between the demand for physical infrastructure and the demand for government regulation such as foreclosure rules. The demand for either may be imputed or based on complaints received. Imputation is based on indicators (see Section 2.2 in Chapter 4) such as overcrowding, heating, plumbing and communication infrastructure, housing affordability, social vulnerability, etc. The result is an inadequate housing map, which may be augmented by point data referencing complaints to a 311 hotline.

Figure 5.22 (San José, 2022; Santa Clara, 2022) shows a mismatch between the imputed and expressed housing demand measures; a discrepancy that is all too common: complaint calls are as much a function of a sense of entitlement or a lack of trust in the efficacy of 311 calls as they are of actual needs. The imputed indicator may hence be better analyzed in light of vulnerable populations such as children, the

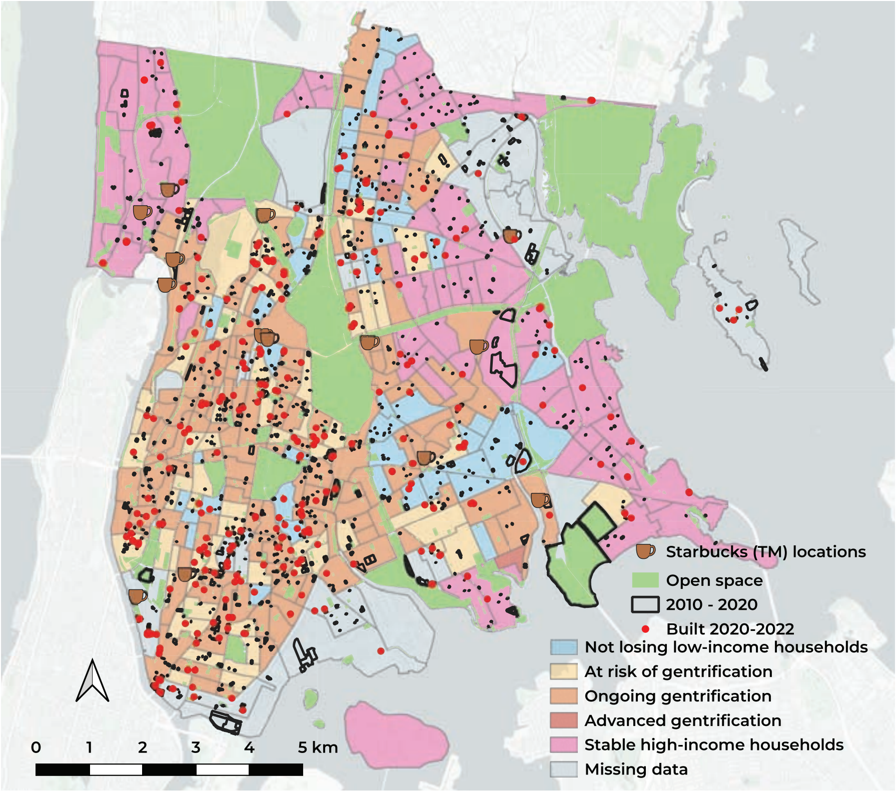


FIGURE 5.20 Bronx stages of gentrification

elderly, or people with disabilities. (See also Section 6.2 in Chapter 6). The map in Figure 5.22 shows that the majority of severe housing shortages lie in a ring around the city center. How would this change if we weigh the absolute number by accessibility to public transit or the provision of medical services? Even at this basic level of evaluation, there are a multitude of GIS operations to be applied – and none of these questions could be addressed by spreadsheets alone.

GIS-supported housing policy evaluations can be distinguished by time or by space. The former is a classic change analysis of, say, an urban revitalization project, while the latter requires the comparison across a spatial boundary separating the study area into parts where the policy is applied as opposed to those where the policy has not changed (e.g., a transit hub on the edge of a municipal jurisdiction). For an evaluation along a temporal axis, the process is similar to the identification of milestones and deliverables in project management. At each stage of the project, inventories are taken and then compared.

Jurisdictional boundaries lend themselves to the planning equivalent of working with control groups in a medical experiment. Many metropolitan areas in the United States have beltways that separate a larger city from its surrounding municipalities. As public transit follows these existing corridors and transit-oriented development fosters densification around transit stations (see Section 3.2.6), these become living laboratories for the effect of different housing policies as they are implemented by

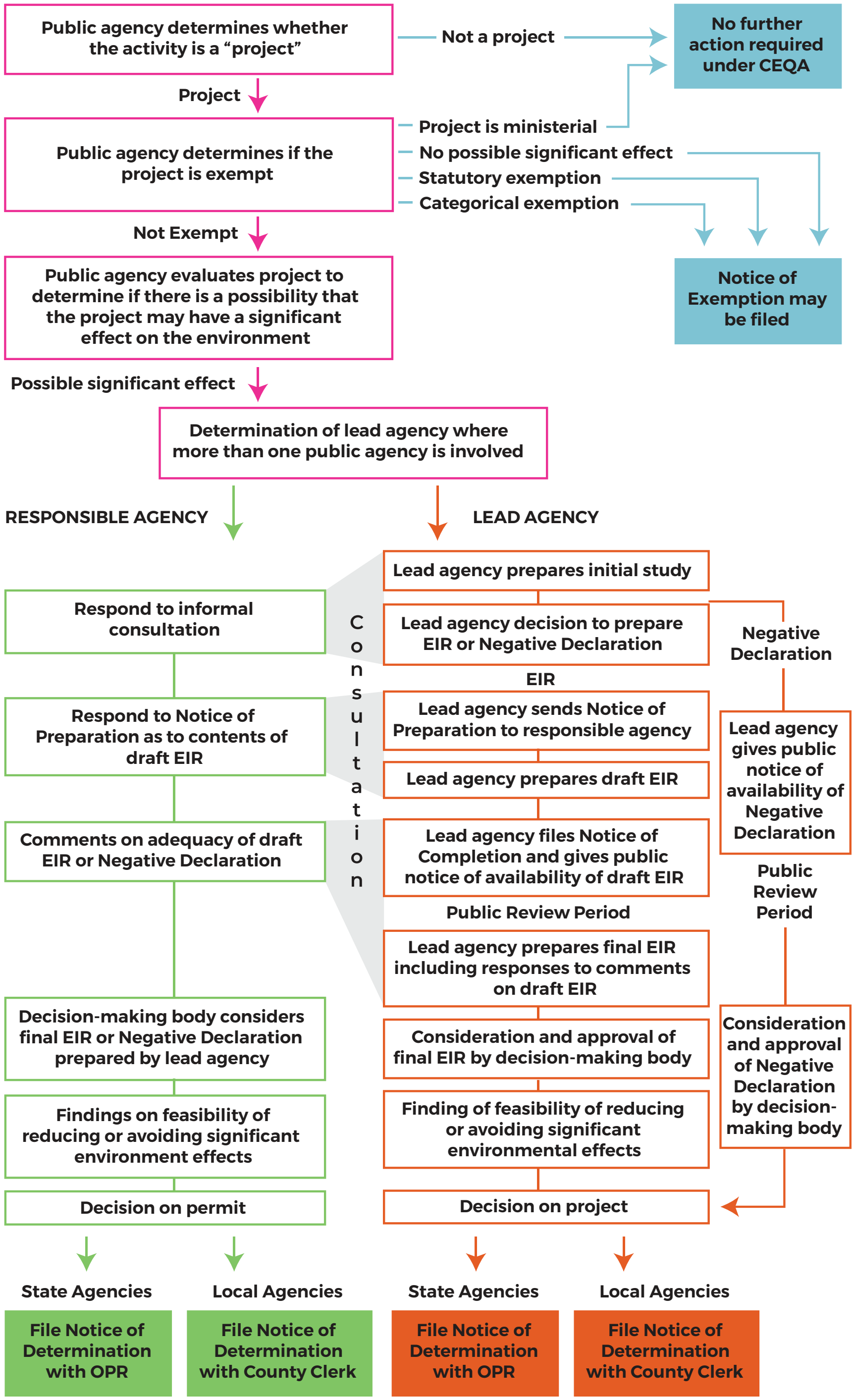

Public agency determines whether the activity is a "project"
Not a project
No further action required under CEQA
Project
Public agency determines if the project is exempt
Project is ministerial
No possible significant effect
Statutory exemption
Categorical exemption
Notice of Exemption may be filed
Not Exempt
Public agency evaluates project to determine if there is a possibility that the project may have a significant effect on the environment
Possible significant effect
Determination of lead agency where more than one public agency is involved
RESPONSIBLE AGENCY
LEAD AGENCY
Lead agency prepares initial study
Respond to informal consultation
Lead agency decision to prepare EIR or Negative Declaration
Negative Declaration
Consultation
EIR
Respond to Notice of Preparation as to contents of draft EIR
Lead agency sends Notice of Preparation to responsible agency
Lead agency gives public notice of availability of Negative Declaration
Lead agency prepares draft EIR
Comments on adequacy of draft EIR or Negative Declaration
Lead agency files Notice of Completion and gives public notice of availability of draft EIR
Public Review Period
Public Review Period
Lead agency prepares final EIR including responses to comments on draft EIR
Decision-making body considers final EIR or Negative Declaration prepared by lead agency
Consideration and approval of final EIR by decision-making body
Consideration and approval of Negative Declaration by decision-making body
Findings on feasibility of reducing or avoiding significant environment effects
Finding of feasibility of reducing or avoiding significant environmental effects
Decision on permit
Decision on project
State Agencies
Local Agencies
State Agencies
Local Agencies
File Notice of Determination with OPR
File Notice of Determination with County Clerk
File Notice of Determination with OPR
File Notice of Determination with County Clerk

FIGURE 5.21 The CEQA flowchart

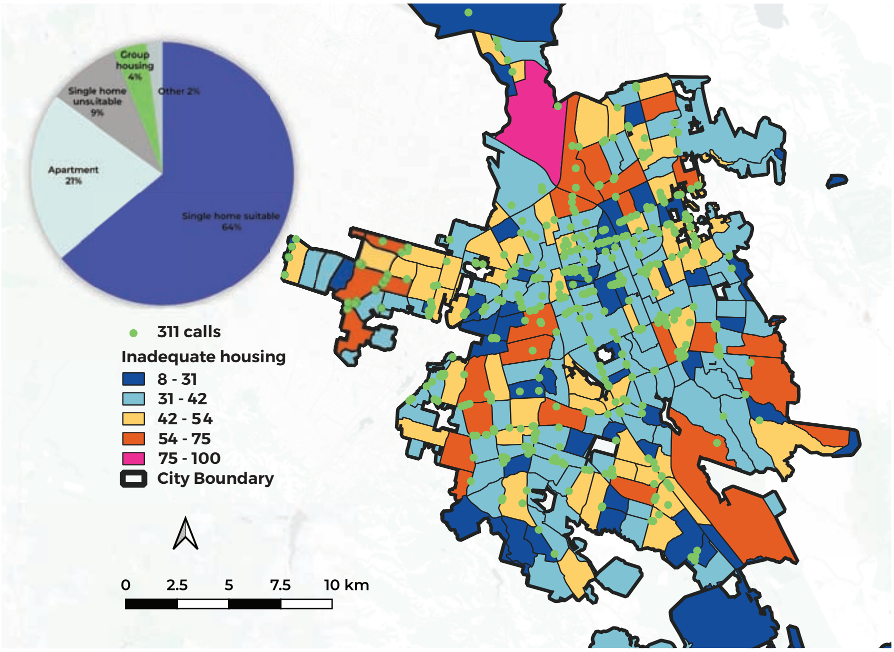


FIGURE 5.22 Tracking community concerns: map and pie chart revealing 311 calls and inadequate housing issues in San Jose, CA

varying local authorities in their vicinity. Auerbach et al. (2020), for instance, report on the use of GIS to compare the effect of an anti-displacement tax fund on West-Atlanta neighborhoods that participate in the effort compared to those who do not.

5.6 ADVANCED TECHNIQUES

The previous two sections provided a pretty thorough introduction to GIS for housing policy research. We laid the technical foundations in Chapter 4 and then delved into the necessary concepts of GIS data models and the main (most commonly used) analysis operations. These sections, in conjunction with a bit of trial and error or learning by doing, will enable diligent readers to use GIS in their everyday housing policy work. The remainder of this chapter is a high-level overview of more advanced GIS techniques available to seasoned housing researchers. This section covers material commonly taught in one or two graduate-level GIS courses but can, of course, not be as thorough. Novices are invited to read this section to learn about topics that may relate to experiences outside the geospatial realm. Readers with some GIS experience will discover applications that go beyond the traditional buffer and overlay paradigm. This section is heavily annotated with links for further readings.

5.6.1 Dasymetric Mapping and Pycnophylactic Interpolation

The term dasymetric mapping (DM) is misleading as it suggests a visualization technique. While it can be used as such, its importance lies mainly in the impact it has

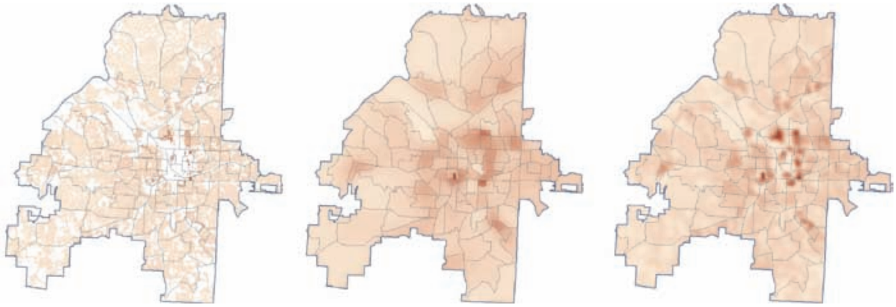


FIGURE 5.23 Atlanta population (a) mapped dasymetrically (b) interpolated pycnophylactically, and (c) with both techniques combined. Based on Kim and Yao (2010)

on analyses. DM is essentially a response to the ecological fallacy of trying to reason about something specific based on only general data. Take US Census tract data for instance. As with all polygon data, it says nothing about how the phenomenon is distributed within a census tract. But we do know that the population the data is based on are residences. And we know that (with very few exceptions) people do not live on water or in parks, parking lots, etc. So, we can redistribute our population-based Census data to those parts of a tract that remain after we have subtracted (another basic GIS operation) the uninhabitable areas. We may even, if we have access to building footprint data, limit the distributions to the building footprints themselves. The result is a much more realistic representation that deals with one aspect of the modifiable area unit problem (MAUP).

The other aspect is that of arbitrarily drawn boundaries that artificially separate continuous phenomena. Returning to our census population example, it is just not reasonable to assume that population characteristics change at the boundary between two tracts. Tobler (1979) developed a technique called *pycnophylactic interpolation* (PI) that takes information about the distribution of a phenomenon in neighboring regions to redistribute the data within each region (e.g., a tract) and create smooth transition across boundaries. The implementation requires translating vector to raster data and having local knowledge about the existence of discontinuities such as rivers, parks, railway lines, etc., all of which would render any interpolation assumption incorrect.

Kim and Yao (2010) present examples that combine dasymetric mapping (DM) and pycnophylactic interpolation (PI) to create data that seems to mysteriously be of much higher accuracy than any input data, see Figure 5.23 (US Census, 2010; Atlanta Regional Commission, 2010). This is reminiscent of the Bayesian approach, where the incorporation of auxiliary data (such as land use) results in much improved interpolation results. It is particularly useful in situations where we try to work with relatively coarse data like from the public health sector. Rather than trying to take our analysis to the parcel level, we can try to improve those coarser datasets so as to not water down our results. In addition to the need to handle the transition between raster and vector data and to find software that performs the pycnophylactic interpolation (R or Python), the main concern is that the processes and results of either DM or PI are consistent with the conceptual model of the researcher. This means that she has to be aware of the assumptions that underlie the creation of the original datasets, in particular its spatial support.[4]

5.6.2 Pattern and Cluster Analysis

We have, by now, presented dozens of maps to illustrate one argument or the other. The built in assumption has been that the map shows the distribution of a particular phenomenon and that the patterns on the map are (i) real, i.e., they can be observed if we visit the place depicted and (ii) are pertinent or (statistically) relevant. The former is difficult to maintain because most of our maps are actually abstractions that have to be translated back into the experiential knowledge of a local observer. The latter takes us into the realm of spatial statistics, which is necessitated by the fact that humans have an uncanny ability to detect patterns where objectively there aren't any (Goldstone and Barsalou 1998; Reber et al. 1998; Rensinck and Baldridge 1998). In other words, we are neurologically hardwired to detect patterns because they are the basis of object recognition and hence our ability to navigate and make sense of the world. This is then, where pattern and cluster analysis come to bear.

The majority of applications are based on point data (e.g., crime locations, 311 calls, grocery stores) because the geometries are easier to run calculations on than with linear or areal features. And here, it is easier dealing just with locations rather than weighing them by some attribute value (e.g., square footage of the grocery store). The question of spatial support raises its head again because something as innocuous as bus stops cannot be randomly distributed as they are spatially constrained by the road network. It is the lack of randomness in urban spatial phenomena that invites spatial statistical analysis. All spatial pattern analyses are about comparing the observed pattern to a set of random patterns to then determine whether the observed one is likely to be random or not. If it could be random (without some chosen confidence interval) then we declare the pattern to not be statistically significant. Matters are complicated by (i) the definition of the boundary of our study area (for instance, we don't expect burglaries to occur inside lakes or water bodies, although theft of fish or water would be another matter entirely) and (ii) the scale of analysis. Something may look like a pattern at one scale but not at another. This, however, points to one of the purposes of the analysis in the first place. Just identifying a pattern is hardly enough; we then want to determine what are the drivers behind the distributions that we observe – and scale dependency helps us to limit the range of possible drivers.

When we determine that our observed pattern is not consistent with randomness, there are two possibilities: the observed pattern may exhibit signs of (i) clustering or of (ii) dispersion. Small amounts of either are normal and would be expected in a random distribution but consistent or strong patterns of clustering or dispersion (e.g., the distribution of black and white fields on a chessboard) point to some forcing factor.

A cluster is described as the intensity of the phenomenon: the more observations in a small area, the more intense the phenomenon (crime, Covid-SARS cases, etc.). This is measured by a so-called kernel density function, where a small (size to be determined and usually the procedure is repeated for many different sizes) search window is continuously moved over the study area to count the number of observations within the search window. The systematic application of varying search window sizes helps with the determination of the pertinent scale of the observed clustering.

The detection of patterns in areal data (e.g., census tracts) requires a discussion of spatial autocorrelation. The same Tobler of pycnophylactic interpolation was coined in an obscure article in 1970 *The First Law of Geography*, which states “everything is related to everything else, but near things are more related than distant things”. It underlies all work in spatial analysis and is the basis for any scientific approach to geography (including GIS) because without it phenomena would be distributed randomly in space and we would have no way to systematically reason about them. Statistically, the first law captures autocorrelation, i.e., the correlation of a variable with itself as a function of distance.[5] The analogy of a chessboard helps again. The black and white fields are perfectly negatively autocorrelated, i.e., every white field shares on all sides boundaries with black fields and vice versa. The position of the figures at the beginning of the game is exactly the opposite: all white figures have only white neighbors and all black figures on black ones. This simple arrangement is harder to discern when the areas are irregular (like Census area units). We then have to establish who is a neighbor of whom (the topological relationships we discussed in Section 5.3 of this chapter), which is encoded in the form of weight matrices that establish the degree of neighborship. There are multiple measures of spatial autocorrelation with the most common one probably being Moran’s *I*, which is a global measure of the relationship between spatial proximity and variable similarity. A local version known as local indicator of spatial association (LISA) captures the difference between the spatial autocorrelation of a small set of neighbors compared with the global measure. It is used to identify so-called hotspots and coldspots (see Figure 5.24 (San José Bikeways, 2022).

5.6.3 Geographically Weighted Regression

> Imagine reading a book on the climate of the United States which contained only data averaged across the whole country, such as mean annual rainfall, mean annual number of hours of sunshine, and so forth. Many would feel rather short-changed with such a lack of detail. We would suspect, quite rightly, that there is a great richness in the underlying data on which these averages have been calculated; we would probably want to see these data, preferably drawn on maps, in order to appreciate the spatial variations in climate that are hidden in the reported averages. Indeed, the averages we have been presented with may be practically useless in telling us anything about climate in any particular part of the United States. It is known, for instance, that parts of the north-western United States receive a great deal more precipitation than parts of the Southwest and that Florida receives more hours of sunshine in a year than New York. In fact, it might be the case that not a single weather station in the country has the characteristics depicted by the mean climatic statistics.

This is the introductory paragraph for *Geographically Weighted Regression* by Fotheringham et al. (2002). And the paragraph describes succinctly one of the main points that we are trying to make in this volume, namely that (i) space/location matters, (ii) that things are not uniformly distributed throughout a region, and (iii) that we have to distinguish between local and global phenomena, where the definition of what constitutes local is variable. This then begs the question how to define a *local* regime or realm of influence. This is exactly what geographically weighted regression (GWR) is good for to answer.

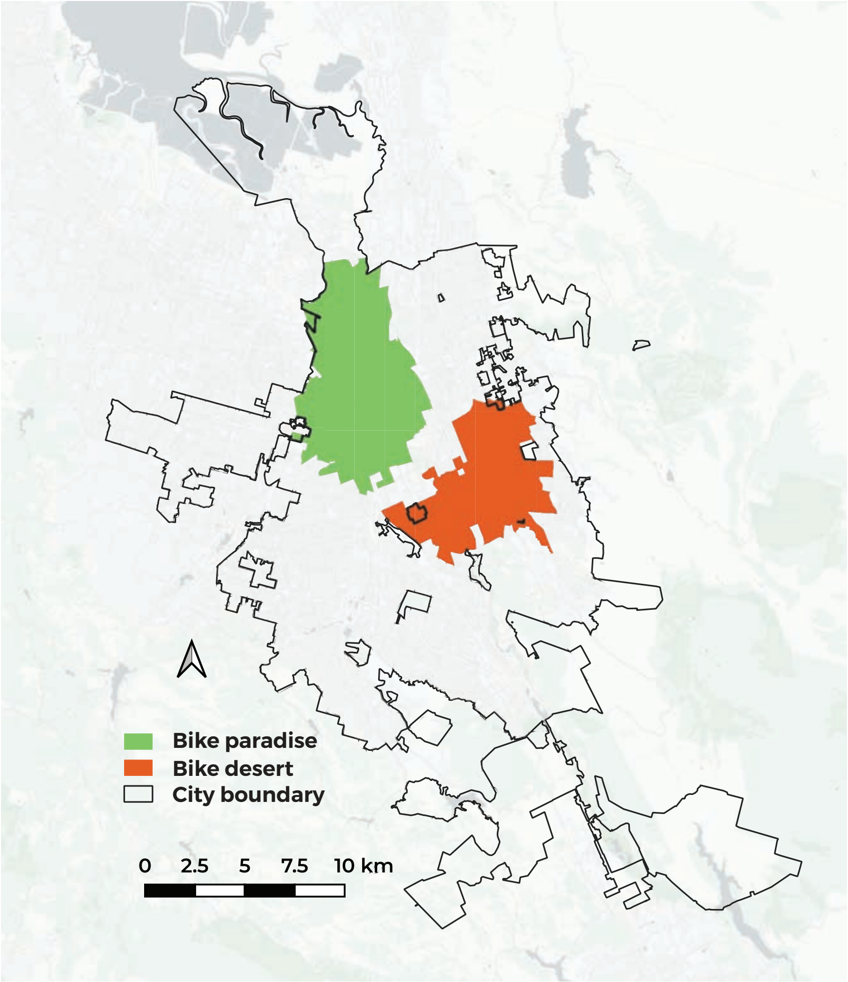


FIGURE 5.24 San Jose bike desert. From Zandiatashbar et al. (2023)

To appreciate the problem that the GWR is trying to solve, let's have a look at a regression model that tries to explain house prices based on a few explanatory variables such as size of the property, amenities, building age, and unemployment rate. A traditional regression model would give us an equation like

$$p = \alpha_0 + \alpha_1 \text{ propsize} + \alpha_2 \text{ amenities} + \alpha_3 \text{ bldage} + \alpha_4 \text{ unemploy} + \varepsilon$$

The error term ε, covering the unexplained component(s) of our model, would then be assumed to be randomly distributed over our study area. As it turns out, however, this is not the case, and it is easily visualized by mapping the difference between the expected and the observed values as in Figure 5.25 (US Census, n.d.; NYC Transit, 2020).

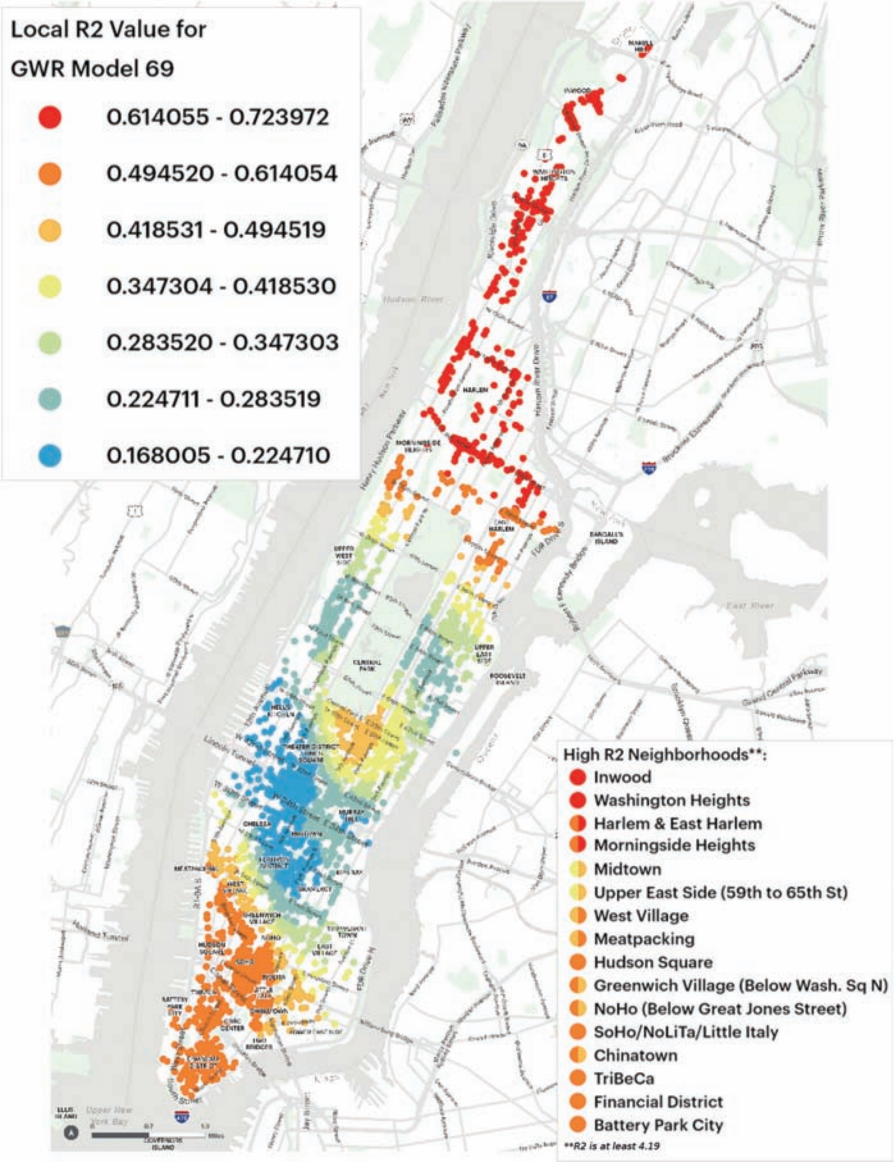


FIGURE 5.25 Non-random (spatially auto-correlated) distribution of residuals in a global regression model

Clearly, when we look at Figure 5.25, we can detect that the residuals are not randomly distributed as we would expect from a random process. We could verify this impression by performing a Moran's *I* spatial autocorrelation test. In other words, the contribution of individual explanatory variables varies over our study area, e.g., the effect of property size on the final price is different in one part of the study area compared to some other. An observant reader might object that this may be due to the MAUP, and if we had chosen area boundaries appropriately, then the map would look very different. However, this is not the case as can be shown if we do not work with polygons but with point data (each individual home sale), which would result in a density map of residuals.

The next logical step would then be to create individual local regression models for each of the ZIP code areas in Figure 5.18. In addition to this getting rather tedious, we would now indeed run into the MAUP, so this is not a practical solution – especially if the footprints for the explanatory variables are varied. The solution comes in the form of a technique adopted from point pattern analysis called moving window regression. A search window of a fraction of the size of the study area is continuously moved over the study and the regression is applied to all the observations that fall within the search window. The MAUP is then resolved by not having the search window jump by the width of its size but say by 1/10th of its size. This smoothes the differences between the regression results and does not assume any boundaries. This is computationally intensive and we would leave it at that if we have a good idea of how far neighborhood effects extend for a particular variable. If this is not the case, then we would run the same GWR procedure with varying window sizes and instead of square windows would employ so-called kernels with varying distance decay functions (Figure 5.26).

The effect of this procedure is three different outcomes, two of which are important, while the third one is contentious. First, when we now map the residuals, we will find that there is no spatial autocorrelation to them and that they are indeed randomly distributed – as we should expect from a regression model. Second, the GWR gives us areas of likewise spatial regimes where the respective regression equations are either the same or very similar. These areas are not the result of any boundaries in the input data but constitute a regionalization of our dependent variable. The importance of this statement is hard to overemphasize; the GWR tells us where, in spite of the curse of spatial variation, we can expect uniform behavior in response to our

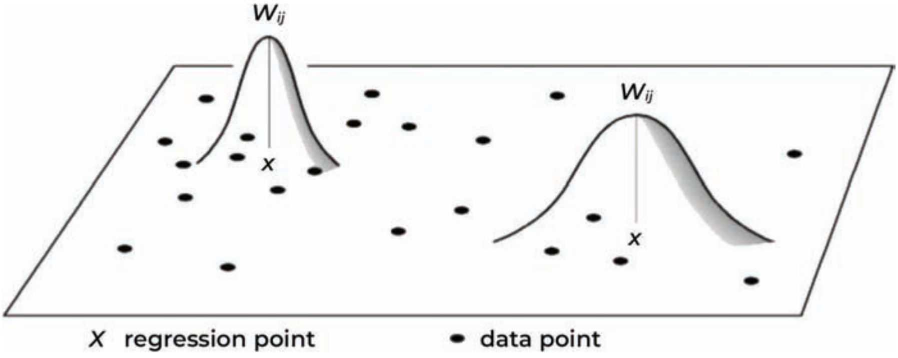


FIGURE 5.26 Varying kernel sizes to emphasize the contribution of neighboring observations as a function of distance

policy decisions. Finally, and this is the contentious part, the GWR gives us usually much improved r^2 values that make us feel good but that many in the community of professional statisticians declare to be unfounded. If the purpose of our analysis is a sound explanatory model, then we have to resort to the spatial regression techniques of the following sub-section. But the value of outcome (2), the regionalization of our research question should not be underestimated.

5.6.4 Spatial Regression

In the most general sense possible, a regression equation describes the relationship between a dependent variable whose value we want to predict on the left-hand side and any number of independent variables that serve to explain the outcome as in this equation:

$$\text{outcome} = \alpha_1 var_1 + \alpha_2 var_2 + \alpha_3 var_n + \varepsilon$$

In non-spatial applications, the parameters α provide a kind of weight (which may also be negative as when higher incomes usually suggest fewer single parents).[6] It is good statistical practice to work with variable values that have been transformed to standardized ranges to ascertain that the parameters relate appropriately to each other. The additional twist in spatial versions of a regression equation is that each α is in turn adjusted by what is known as a spatial weight matrix. The spatial weight matrix is a construction that specifies the influence that the value of one observation has on its neighbors and is usually distance-weighted, i.e., observations further away have a lesser influence (Tobler's First Law). There are a multitude of methods to create such a spatial weight matrix, depending on the type of geometry as well as how many neighbors should be incorporated and the reader is referred to standard textbooks such as LeSage and Pace (2009) and Anselin and Rey (2010), or Anselin and Rey (2014).

The obvious reason for the construction of the spatial weight matrix is to deal with spatial autocorrelation; something that is seen as a nuisance in traditional statistics but is now employed as an additional piece of information. The GWR from the previous subsection implicitly creates an optimized spatial weight matrix but does not export it for further exploration or comparison. In another twist, spatial influences may not just impact the values of each explanatory variable but may also be hidden in the error term ε. Models addressing the former are referred to as spatial lag models (explaining the influence that neighbors have), while the latter is known as spatial error models. The spatial lag $y_{lag\text{-}i}$ is

$$y_{lag-i} = \sum_j w_{ij} y_j$$

where $y_{lag\text{-}i}$ is the spatial lag of variable y at location i, and j sums over the entire dataset. For spatial error models the traditional e is replaced with $u_{lag\text{-}i} + \varepsilon_i$.

Traditional GIS are not made for this but many of the bigger statistics programs have modules for spatial regression; none more so than the statistics package R.

5.7 FURTHER READING

These last two chapters concentrated on the technical aspects of GIS for housing policy research. Readers who want to go beyond what has been presented here will want to peruse some of the readings suggested in the next free paragraphs. However, before you take off to another book, let's have a look at what the next chapter has to offer.

While Chapters 1–3 provided an overview of the housing policy landscape and the kinds of problems we are trying to solve, Chapter 4 introduced us to the geospatial data that then allows us to make use of the unique capabilities of GIS in Chapter 5. One of the tenets of this book is that the geographic perspective of spatial differentiation has been underdeveloped in much of the housing policy literature. Many problems can only be addressed if they are seen both in concert with the perspectives of related fields as well as the unique set of circumstances/conditions that makes each location unique. With this in mind, we can now apply the GIS tools introduced here to the big challenges that every housing researcher is confronted with in the 21st century. Regardless of whether we want to overcome the single-family residential paradigm, modernize housing and neighborhood design, deal with the changes of mobility patterns brought about by the diversification and hybridization of work, combat homelessness and housing insecurity, deal with climate change, public health or public safety, GIS lies at the center of each solution space. In Chapter 6, we will illustrate through numerous examples how GIS is used to address each of these challenges.

5.7.1 GIS Models

A good overview of vector data formats can be found in Diamond (2019), while the corresponding article for raster formats is Williams (2019). Conceptual data models, including tools and languages to compile them, are well covered in Nyerges (2017a). From a GIS project development perspective, this should precede the choice of logical data model described by the same author in (2017b).

A very brief introduction to conceptual ways of organizing spatial data is Varanka's (2021) article, however, the reader might want to skip right down to the end of this encyclopedia entry to find truly further readings; it lists many classics that should be on the shelf of every GIS practitioner. Two specific data models discussed in our volume are the raster and the vector model. A nice overview of the former is Pingel (2018), which is complemented by Albrecht's (2022) discussion of entity-based models. Albrecht's article also makes for a good entry point to the next section on basic GIS analysis operations.

5.7.2 Basic GIS Analysis Operations

Spatial neighborhoods can be defined in many different ways and Mu and Holloway (2019) provide a nice overview. Interestingly, they miss a crucial body of work epitomized by the Laval school of geomatics. Gold's (2016) article on tessellations would be a good representative of that line of thinking. Another fundamental approach to understanding basic GIS analyses is set theory. Arlinghaus' (2019) article is a good starting point. This leads directly to overlay analysis as introduced by Cai (2022), the counterpart to which would be Li's (2017) entry on buffering.

5.7.3 Advanced GIS Techniques

A good introduction to dasymetric mapping is Mennis' (2017) encyclopedia entry. It builds on Tobler's (1979) article on pycnophylactic interpolation, which is eminently readable in spite of its publication in the Journal of the American Statistical Association. Pattern and spatial cluster analysis are common techniques in landscape ecology and crime analysis. There are thousands of applications but the original description in McGarigal and Marks (1995) remains the go-to reading on this subject.

An excellent user-friendly introduction to a range of spatial (statistical) analysis techniques is the GeoDa software developed by the Center for Spatial Data Science at the University of Chicago. GeoDa incorporates a range of spatial analysis methods in a very user friendly way, one of which is Local Indicators of Spatial Association, first described by Anselin (1995). One technique not covered by GeoDa is Geographically Weighted Regression (GWR), epitomized by Fotheringham et al. (2003). Although eminently readable, readers of this volume might want to start with Sachdeva and Fotheringham's (2020) overview. Chakraborty and McMillan's (2022) article entitled "Is Housing Diversity Good for Community Stability?" is a nice example of the application of *spatial regression* in housing research.

NOTES

1. The same problem occurs in the world of mankind as well; see, for example, the ill-defined boundaries of neighborhoods or regions such as the boundary between the eastern United States and the Midwest.
2. The American meteorologist Alfred Thiessen (1911) and the Ukrainian mathematician Georgy Voronoi (1908) introduced these structures to a geophysical community at roughly the same time without knowing about the respective other's work. They were both preceded by the German mathematician Dirichlet, who in 1850 in his *Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen* defined what in mathematics is known as Dirichlet regions.
3. Database aficionados would beg to differ as all of this can also be done with spatial SQL.
4. In mathematics, the support of a real-valued function f is the subset of the domain containing the elements which are not mapped to zero. If the domain of f is a topological space, the support of f is instead defined as the smallest closed set containing all points not mapped to zero.
5. Outside of geospatial applications, auto-correlation is typically understood to be the correlation of a variable with itself as a function of a lag or distance in time.
6. A negative variable weight α_n indicates that the outcome increases as the variable value decreases. If, for example, the outcome variable is median area income, then a smaller number of single parents typically results in a higher area income (and vice versa).

REFERENCES

Albrecht, J, 2022. "Entity-Based Models". In: John P. Wilson (Ed.), *Geographic Information Science & Technology Body of Knowledge*. doi:10.22224/gistbok/2022.2.11.

Anderson, J, 2015. *Public Policymaking*. Stamford, CT: Cengage Learning.

Anselin, L, 1995. "Local Indicators of Spatial Association-LISA". *Geographical Analysis*, 27: 93–115, doi:10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L, and Rey, S, 2010. *Perspectives on Spatial Data Analysis*. Berlin/Heidelberg: Springer. doi:10.1007/978-3-642-01976-0.

Anselin, L, and Rey, S, 2014. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. Chicago, IL: GeoDa Press LLC.

Arlinghaus, S, 2019. "Set Theory". In: Wilson, J (Ed.), In: *The Geographic Information Science & Technology Body of Knowledge* (2nd quarter, 2019 edition). doi:10.22224/gistbok/2019.2.1.

Association of Environmental Professionals, 2022. *2022 CEQA California Environmental Quality Act Statute and Guidelines*. Palm Desert: California AEP.

Baltimore, DHCD. *CoDe Map*, n.d. https://cels.baltimorehousing.org/codemapv2ext/, 14 July 2023.

Bucks County Planning Commission, n.d. *Proposed Subdivisions and Land Developments*. https://bucksgis.maps.arcgis.com/apps/webappviewer/index.html?id=f58e99f72c4241ebbe309e08d6e42198, 14 July 2023.

Cai, H, 2022. "Overlay". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (1st quarter 2022 edition). doi:10.22224/gistbok/2022.1.2.

Chava, J, and Renne, J, 2022. "Transit-Induced Gentrification or Vice Versa?" *Journal of the American Planning Association*, 88(1): 44–54.

Checker, M, 2011. "Wiped Out by the 'Greenwave': Environmental Gentrification and the Paradoxical Politics of Urban Sustainability". *City & Society*, 23(2): 210–29. doi:10.1111/j.1548-744X.2011.01063.x

Chu, M, Fenelon, A, Rodriguez, J, Zota, and Adamkiewicz, G, 2022. "Development of a Multidimensional Housing and Environmental Quality Index (HEQI): Application to the American Housing Survey". *Environmental Health*, 21: 56. doi:10.1186/s12940-022-00866-8.

Chester County Planning Commission, n.d. *Chester County Planning Commission*. https://www.chescoplanning.org/planreview/Maps.cfm, 14 July 2023.

Devillers, R and Jeansoulin, R, 2006. "Fundamentals of Spatial Data Quality". London: Wiley. doi:10.1002/9780470612156.

Diamond, L, 2019. "Vector Formats and Sources". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (4th quarter, 2019 edition), https://doi/10.22224/gistbok/2019.4.8.

Donnelly, F, 2022. *US Census Data: Concepts and Applications for Supporting Research*. American Library Association Library Technology Reports (vol. 58, no. 4). Chicago, IL: ALA TechSource.

FEMA, 2022. "Hazus 6.0 Baseline Data Updates". *FEMA Factsheet*. https://www.fema.gov/sites/default/files/documents/fema_hazus-6-data-updates-factsheet.pdf, last accessed 4 December 2022.

Fotheringham, A, Brunsdon, C, and Charlton, M, 2003. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. New York: John Wiley & Sons.

Geo, DA, 2023. https://geodacenter.github.io/, last accessed 23 May 2023.

Gold, C, 2016. "Tessellations in GIS: Part I-putting it all together". *Geo-Spatial Information Science*, 19(1): 9–25, doi:10.1080/10095020.2016.1146440

Goldstone, R, and Barsalou, L, 1998. "Reuniting Perception and Cognition". *Cognition*, 65: 231–262.

Goodchild, M, Li, W, and Tong, D, 2022. "Introduction to the Special Issue on Scale and Spatial Analytics". *Journal of Geographical Systems*, 24: 285–289. doi:10.1007/s10109-022-00391-9

Guptill, S, and Morrison, J, 1995. *Elements of Spatial Data Quality*. Amsterdam:Elsevier.

Hirschman, C, Alba, R, and Farley, R, 2000. "The Meaning and Measurement of Race in the U.S. Census: Glimpses into the Future". *Demography*, 37: 381–393.

HUD PD&R Edge, 2017. *Defining Housing Affordability*. PD&R Edge. Washington, DC: U.S. Department of Housing and Urban Development's (HUD's) Office of Policy Development and Research (PD&R). https://www.huduser.gov/portal/pdredge/pdr-edge-featd-article-081417.html, last accessed 04 March 2023.

Jacobs, K, and Mitchell, S, 2020. "OpenStreetMap Quality Assessment Using Unsupervised Machine Learning Methods". *Transactions in GIS*, 24: 1280–1298. doi:10.1111/tgis.12680.

LeSage, J, and Pace, R, 2009. "Introduction to Spatial Econometrics". New York: Chapman and Hall/CRC. doi:10.1201/9781420064254.

Li, X, 2017. "Buffers". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (4th quarter, 2017 edition). doi:10.22224/gistbok/2017.4.10.

MapPLUTO, n.d. *New York City Department of Urban Planning and Development parcel-level database*. Online resource available at https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.

McGarigal, K, and Marks, B, 1995. *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure*. USDA Forest Service General Technical Report PNW-351, Corvallis. https://www.fs.usda.gov/pnw/pubs/pnw_gtr351.pdf, last accessed 22 December 2022.

Mennis, J, 2017. "Dasymetric Mapping". In: Richardson, D, Castree, N, Goodchild, M, Kobayashi, A, Liu, W, and Marston, R (Eds.), *International Encyclopedia of Geography: People*, *the Earth*, *Environment and Technology*. Hoboken, NJ: Wiley. doi:10.1002/9781118786352.wbieg0443.

Missouri Census Data Center, 2022. "Intro to Census Geography, Summary Levels, and GeoIDs". https://mcdc.missouri.edu/geography/sumlevs/, last accessed 4 December 2022.

Mu, L, and Holloway, S, 2019. "Neighborhoods". In: Wilson, J (Ed.), *The Geographic Information Science and Technology Body of Knowledge* (1st quarter, 2019 edition). doi:10.22224/gistbok/2019.1.11.

National Institute of Standards and Technology (NIST), 2021. *Compliance FAQs: Federal Information Processing Standards (FIPS)*. https://www.nist.gov/standardsgov/compliance-faqs-federal-information-processing-standards-fips, last accessed 4 December 2022.

NYC Open Data, n.d. NYC Open Datamine, an online resource available at https://opendata.cityofnewyork.us/data/.

NYC Transit, 2020. *Turnstyle passenger counts at New York City subway stations*. Online resource available at https://qri.cloud/nyc-transit-data/turnstile_daily_counts_2020.

Nyerges, T, 2017a. "Logical Data Models". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (1st quarter, 2017 edition), doi:10.22224/gistbok/2017.1.2.

Nyerges, T, 2017b. "Conceptual Data Models". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (1st quarter, 2017 edition). doi:10.22224/gistbok/2017.1.3.

OGC, 2023. *E-Learning Documentation*. https://opengeospatial.github.io/e-learning/geopackage/text/contents.html, last accessed 24 May 2023.

Openshaw, S, and Taylor, P, 1979. "A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem". In: Wrigley, N (Ed.), *Statistical Methods in the Spatial Sciences*, pp. 127–144. London: Pion.

OPR, 2023. *CEQA: The California Environmental Quality Act*. Sacramento, CA: Governor's Office of Planning and Research. https://opr.ca.gov/ceqa/, last accessed 30 May 2023.

Pingel, T, 2018. "The Raster Data Model". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (3rd quarter, 2018 edition). doi:10.22224/gistbok/2018.3.11.

R Spatial, 2023. https://cran.r-project.org/web/views/Spatial.html, last accessed 30 May 2023.

Reber, R, Schwarz, N, and Winkielman, P, 1998. "Effects of Processing Fluency on Affective Judgments". *Psychological Science*, 9: 45–48.

Rensink, R, and Baldridge, G, 2010. "The Perception of Correlation in Scatterplots", *Computer Graphics Forum*, 29(10): 1203–1210.

Sachdeva, M, and Fotheringham, A, 2020. "The Geographically Weighted Regression Framework". In: Wilson, J. (Ed.), *The Geographic Information Science and Technology Body of Knowledge* (4th quarter, 2020 edition), doi:10.22224/gistbok/2020.4.7.

San José, 2022. *Database of 311 calls to the San Jose, CA, call center*. Online resource available at https://311.sanjoseca.gov/.

San José Bikeways, 2022. *Bicycle network dataset*. Online resource available at https://data.sanjoseca.gov/dataset/bikeways.

Santa Clara, 2022. *Database of housing quality standards inspections conducted by the Housing Authority of Santa Clara County*, available at schousingauthority.org.

Sparx Systems, 2022. *Guide to Business Modeling*. https://sparxsystems.com/resources/user-guides/16.0/guidebooks/business-modeling-techniques.pdf, last accessed 12/04/2022.

Tobler, W, 1979. "Smooth Pycnophylactic Interpolation for Geographical Regions". *Journal of the American Statistical Association*, 74(367): 519–530. doi:10.2307/2286968

US Census, n.d. *Generic US Census data*. Online resource available at https://data,census.gov.

US Census, 2010. *Block-level total population 2010*. Online resource available at https://data.census.gov.

Varanka, D, 2021. "Data Properties". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (1st quarter, 2021 edition). doi:10.22224/gistbok/2021.1.15.

Westlaw, 2023. *Guidelines of the CEQA*, *in the California Code of Regulations*, *Title 14*, *Natural Resources*. Division 6 Resources Agency. https://govt.westlaw.com/calregs/Index?transitionType=Default&contextData=%28sc.Default%29, last accessed 7 May 2023.

Williams, C, 2019. "Raster Formats and Sources". In: Wilson, J (Ed.), *The Geographic Information Science & Technology Body of Knowledge* (4th quarter, 2019 edition). doi:10.22224/gistbok/2019.4.11

Zandiatashbar, A, Albrecht, J, and Nixon, H, 2023. *A Bike System for All in Silicon Valley: Equity Assessment of Bike Infrastructure in San José, CA*. San José, CA: San José State University, College of Business, Mineta Transportation Institute.