DISCUSSION PAPER SERIES

Understanding Urban Economies, Land Use, and Social Dynamics in the City: Big Data and Measurement

Albert Saiz
Arianna Salazar-Miranda

Understanding Urban Economies, Land Use, and Social Dynamics in the City: Big Data and Measurement

Albert Saiz
MIT and IZA

Arianna Salazar-Miranda
MIT and University of Chicago

ABSTRACT

Understanding Urban Economies, Land Use, and Social Dynamics in the City: Big Data and Measurement*

Recent advancements in data collection have expanded the tools and information available for urban and spatial-based research. This paper presents an overview of spatial big data sources used in urban science and urban economics, with the goal of directing and enriching future research by other applied economists. We structure our discussion around data origins and analytical methods, discussing geographic information maps, GPS and CDR, textual repositories, social media, credit card transactions, street imagery, sensor readings, volumetric data, street patterns, transportation metrics, public records, geocoded historical data, business analytics, real estate transactions, and crowdsourced input. While aiming to provide an overarching perspective, we also touch upon common challenges in urban big data research, especially those unique to data collection, analysis, and inference.

Corresponding author:
Albert Saiz
Department of Urban Studies and Planning
Massachusetts Institute of Technology
77 Massachusetts Avenue, Office 9-332
Cambridge, MA 02139
USA

E-mail: saiz@mit.edu

1 INTRODUCTION

Over the past two decades, we have witnessed an explosion in the availability of novel data sources to measure and understand human behavior in cities. This boom is due to the rapid growth in computer power, connectivity, and data storage capabilities. Such advancements have enabled us to harness new data and utilize the existing sources more efficiently. A growing information technology infrastructure allows for novel insights into urban phenomena across several dimensions: qualitative, quantitative, spatial, and temporal (Ben-Joseph, 2011; Kitchin, 2014).

In urban planning and municipal policy, the "smart city" movement aims to bring revolutionary changes to local government operations (Batty, 2013a; Ratti, 2013; Geertman et al., 2015). At the same time, the private sector has introduced new business applications and investment strategies, capitalizing on large urban datasets (Saiz and Salazar-Miranda, 2017; Barkham et al., 2022).

In the research domain—our main focus in this paper—the new field of "urban informatics" (Foth, 2008; Batty, 2013b; Thakuriah et al., 2017) has taken shape, driven by innovative data sources. However, much of the research still has its roots in traditional disciplines known for their quantitative, mathematical, and statistical models, such as transportation research, computer science, civil engineering, computational design, and urban economics.

Rather than reviewing this work through a disciplinary lens, this chapter highlights various big data sources used in urban research. Our goal is to inspire future research rather than to furnish a catalog of past uses. To this end, we present a number of relevant examples that illustrate each data type. While these examples cover just a fraction of the available literature, they effectively show the types of potential data projects that urban-focused researchers can pursue. We primarily focus on recent research that we are familiar with; thus, our examples predominantly come from urban economics, real estate, and city planning.

When categorizing these data sources, we consider both their origin and collection methodologies. This means that some data sources may span multiple categories. We accept this overlap for the sake of practicality: we aim to cover as much ground as possible and offer the readers a wide breadth of ideas.

We also touch on some of the challenges of using big data in urban research. Instead of delving deep into methodological details, we aim to make readers aware of the potential caveats so they can pursue their study further, either through the references in the paper or their own inquiries.

This paper can be used as a resource for researchers in labor economics or broadly by

those new to urban data. Applied economists in many fields—including Development, Public, Health, and Environmental Economics—routinely use spatial data sets. Labor Economics, specifically, shares a large boundary of research interests with Urban Economics, in topics such as ethnic segregation, human capital externalities, social ties, commuting, work-from-home, labor migration, wage differentials across cities, agglomeration economies, and technological innovation, just to name a few. Our hope is that learning about the use of spatial datasets in other fields can inform and inspire future research along these lines.

The paper proceeds as follows. Section 2 delves into geographic information maps, with a focus on raster, polygon, and point data. Section 3 explores emerging technologies, highlighting the roles of the global positioning system (GPS) and call detail records (CDRs). In Section 4, we examine the utility of text mining techniques for analyzing diverse urban data resources. Section 5 considers both textual and non-textual aspects of social media data, while Section 6 introduces the analysis of credit card transaction data. Section 7 discusses how street imagery can be utilized to study the physical characteristics of cities. Section 8 introduces the deployment of street sensors to capture environmental factors like air pollution and noise. Section 9 is dedicated to exploring building volumes and street shapes, whereas section 10 explores the use of public transportation data. Section 11 features historical data. Section 12 reviews governmental data sources that have been extensively explored in existing literature. Sections 13 and 14 shift the focus to private firm repositories and real estate data, respectively. Section 15 discusses crowdsourced datasets, such as user-generated images. Finally, Section 16 outlines potential challenges and pitfalls associated with urban data collection, analysis, and inference.

2   Data from Geographic Information Systems Maps

Geographic information system (GIS) data and software map the boundaries of spatial objects and also link spatial coordinates to quantitative variables. Early GIS applications, like the Canada Land Inventory in the 1950s and 1960s, primarily focused on mapping boundaries and geographic measurement (Goodchild, 2023). Today, GIS datasets are produced by thousands of providers and accessed by millions worldwide. Berry et al. (2023) note there are over 120,000 studies using GIS methods or data. Nonetheless, tasks such as mapping, visualization, and topographic description remain central to GIS, in sectors spanning from the environment to military technologies.

Oftentimes, quantitative urban researchers extract GIS information to form relational datasets, connecting observations through geographic IDs, such as ZIP codes. Statistical

techniques and modeling are then used to study the relationships between variables in a specified geographic area.

There are many types of GIS data, but some of the most prominent ones are illustrated in Figure 1 and detailed as follows. Raster data operate on a grid system. Each grid cell can be defined by four corner coordinates or by its center point paired with a standard width and length. These cells are associated with various data points, such as differing measurements of the average land slope or multiple readings related to the color that the soil reflects at different wavelengths. Some raster datasets stem from satellite observations capturing the land's topographical or visual attributes. Over the years, the precision of raster data has improved, and it is now available at much higher spatial resolutions (e.g., 30-by-30 meters rather than one-kilometer-square areas).

A Raster   **B** Lines and Points   **C** Polygons
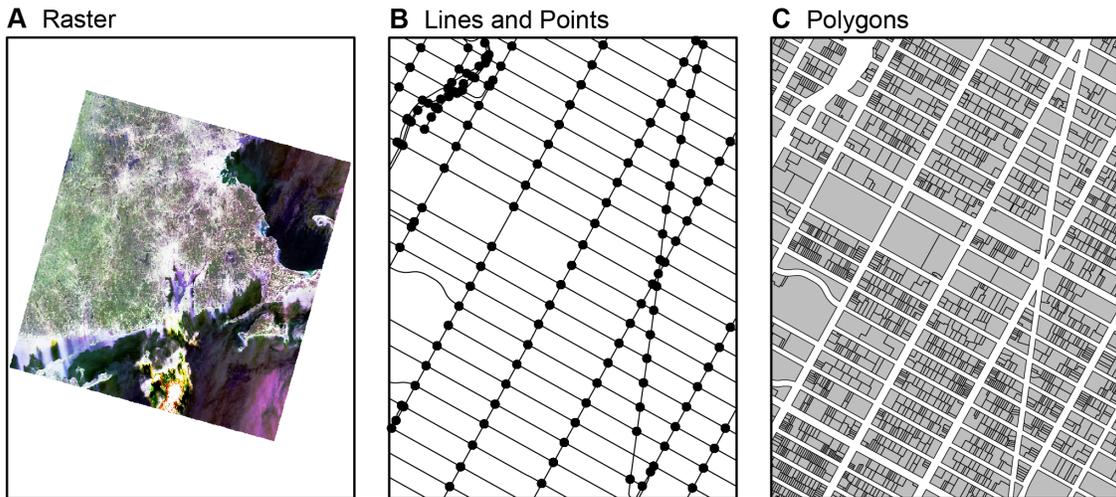
FIGURE 1: Panel A shows raster data obtained from Landsat-7 image courtesy of the US Geological Survey. Panel B shows line and point data obtained from OpenStreetMaps. Panel C shows polygon data obtained from Regrid.

GIS polygons are used to represent administrative or geographic areas such as municipal boundaries or climatic zones. Instead of being defined by a continuous line with infinite points, they are delimited by a substantial but finite set of ordered latitude and longitude points—these are the polygon's vertices. The area enclosed by these vertices corresponds to an approximation of the official boundaries. Data relevant to specific administrative regions or other demarcations can be associated with each polygon. One of the advantages of this approach is the capability to combine or overlay data by uniting or intersecting polygons based on different boundary definitions.

GIS polylines connect a sequence of points to approximate the paths followed by trails, roads, rivers, railways, and other similar features. In contrast, points of interest (POIs) are simply defined by a specific latitude and longitude and often include an ID and a descriptor indicating the type of place, such as a church or school.

2.1 Raster Data: Satellite and Nighttime Imagery

Satellite imagery offers a bird's-eye view of the built environment, enabling detailed assessments of topographical features, land use patterns, urban sprawl, building density, economic activity, and vegetation cover (Donaldson and Storeygard, 2016). By analyzing the spectral characteristics of images, geographic characteristics can be discerned and studied on an expansive scale.

For instance, digital elevation models generated from satellite data can accurately represent the Earth's surface. Saiz (2010) processed satellite data to reveal that steep-sloped terrain significantly limits residential development in urban areas by reducing the available area for construction. Other geographical attributes, such as proximity to green spaces, have been shown to directly impact real estate values (Bolitzer and Netusil, 2000; Anderson and West, 2006; Lutzenhiser and Netusil, 2001; Teo et al., 2023).

Other studies have used satellite imagery to analyze the spatial structure of cities (Kii et al., 2023). For instance, Harari (2020) utilized satellite imagery and historical maps to study the economic implications of city shapes in India, uncovering a direct link between more compact city forms and accelerated population growth.

Furthermore, satellite imagery is helping to measure human activity in cities, population growth, and environmental impacts (Elvidge et al., 2007; Doll et al., 2000). The brightness levels of nighttime imagery, indicative of human presence and activity, have been used to estimate density, urbanization, and economic growth (Ebener et al., 2005; Sutton et al., 2001; Chen and Nordhaus, 2011; Henderson et al., 2012). High activity levels could suggest thriving retail or residential markets, while sudden changes over short periods might point to urban sprawl or gentrification (Sutton, 2003; Lin et al., 2021). Research has harnessed satellite imagery to study urban area growth patterns (Fragkias and Seto, 2007), land use change (Seto and Fragkias, 2005), and the impact of land use fragmentation on municipal expenditures (Rolheiser and Dai, 2019). The objectives of other studies range from assessing infrastructure provision impacts (Zhou et al., 2022) and detecting informal settlements (Wurm et al., 2019; Prabhu et al., 2021) to exploring environmental dimensions such as biodiversity and carbon footprint (Seto et al., 2011, 2012; Mahtta et al., 2019; Burke et al., 2021).

2.2 GIS Polygon Data

One of the most common uses of polygon data is in the analysis of census data. Nowadays, many countries map out their census districts—for instance, census tracts in the US—and offer decennial or annual sample data in GIS format. Some research makes explicit use of the micro-spatial structure of these data. For instance, Saiz and Wachter (2011) use a spatial diffusion model to predict the distribution of immigrants by census tract in the US. Tracts near areas with a high concentration of a particular nationality of immigrants are found to receive more individuals from that group in the next census year. This spatial diffusion process, which resembles the spread models used in epidemiology, accelerates more rapidly in metropolitan areas with higher influxes of immigrants. Moraga et al. (2019) support these conclusions using data from Spain. By explicitly modeling the dynamics of demographic shifts across neighboring census tracts, researchers can isolate quasi-exogenous variation and forecast future changes in the local socioeconomic composition.

Noonan (2005) integrates GIS census tract boundaries with polylines representing major highways and railways. This study shows that these man-made barriers contribute to the segregation of the African American population. Building on this, Ananat (2011) uses the existence of such barriers as a source of quasi-exogenous variation to demonstrate that segregation negatively impacts the African American population.

Urban researchers in the social sciences have also used GIS polygons detailing soil attributes. For instance, Burchfield et al. (2006) measure the distance from metropolitan land plots to the nearest aquifer and show that this variable is a local predictor of urban sprawl in the American West. Rosenthal and Strange (2008) and Barr et al. (2011) utilize an overlay of soil data and ZIP code boundaries, showing that areas situated on bedrock tend to have taller commercial and residential buildings and consequently concentrate more employment and population.

Similar to satellite data, integrating GIS polygon data has been a rich area of exploration in the literature of hedonic pricing in urban economics and real estate (Sirmans et al., 2005; Herath and Maier, 2010). In the typical application, these studies link a housing unit's sale price to its ZIP code or census tract socioeconomic and environmental characteristics. Often, the main objectives are property valuation assessment and forecasting.

On occasion, hedonic methods can shed light on the societal importance of policies or public goods. For instance, Black (1999) overlays residential property sales data—using homes' geo-coordinates—on school attendance boundaries. She then contrasts the sale prices of *adjacent* houses within the same municipality but assigned to different schools. She finds a significant price premium for the homes in better school attendance areas. This approach is

often termed a boundary discontinuity design because it compares properties near the same location but on different sides of an administrative boundary (Keele and Titiunik, 2015). Because the locations are almost identical, and families pay the same tax rates, any real estate price discrepancy can likely be attributed to policies that change discontinuously at the boundary. However, in some circumstances, this research design can be complicated by the existence of spatial spillovers across the boundary line (Jardim et al., 2022).

Sometimes, researchers transform GIS polygon data into point data by calculating the latitude and longitude of a polygon's central point, known as the centroid. This allows for the data to be stored in conventional spreadsheet formats. Often, additional approximations— such as the haversine distance—are used to calculate matrices of distances between centroids. As an illustration, Boarnet et al. (2005) use these distance matrices to analyze the connection between local employment and population growth in Orange County, California.

2.3 POIs and GIS Polylines

Research often quantifies the distance from each geographic observation to a specific POI, such as a landmark or feature. For instance, Müller et al. (2010) and Garcia-López et al. (2015) consider the proximity of geographic areas to highway exits as a factor influencing urban growth. Alternatively to measuring distance, buffer zones can be designated around specific POIs. As shown by Linden and Rockoff (2008), properties within approximately 0.1 mile of an address recently occupied by a registered sex offender see a decline in real estate value. Regarding architectural value, Ahlfeldt and Mastro (2012) find that properties within a 50-to-100-meter radius of the nearest Frank Lloyd Wright building in Oak Park, Illinois, see an 8.5 percent premium price. This premium is about 5 percent for properties within a 0-to-50-meter radius.

POIs are often combined with polygon features, like ZIP codes, to count the number of occurrences within each polygon. The so-called coincidence analyses relate outcomes at the polygon level to the tally of POIs contained within them (Perry and Dixon, 2002). Notably, count data frequently yield high instances of zero observations. For a more continuous variable approach, gravity equation measures of POI accessibility can be crafted. In this type of measure, each observation unit is assigned an "accessibility" score, A, determined by the sum of inverse squared distances to all POIs in the area, represented by $A = \sum_{i=1}^{n} \frac{1}{d_i^2}$. This methodology can be seen in works by Carlino and Saiz (2019) and Heroy et al. (2023). The former paper focuses on POIs such as restaurants and tourism offices, as well as geolocated images from platforms like Flickr and Panoramio—the latter now incorporated into Google Maps. Their findings show that American neighborhoods with greater amenity accessibility

in 2000 witnessed larger economic and demographic growth. Meanwhile, Heroy et al. (2023) find that increased amenity access increases pedestrian activity in Bogota, Colombia. Many other papers use gravity-based accessibility measures, as summarized by Nyerges et al. (2011).

Polylines have a wide array of applications. Occasionally, researchers overlay a set of polygon features onto a polyline. Each polygon then obtains an additional binary variable, assigned a value of one if intersected by the polyline and zero otherwise. For instance, Baum-Snow (2007) demonstrates that central cities in the US that were bisected by a national highway, experienced faster rates of suburbanization between 1950 and 1990.

Another approach involves measuring the proximity to the nearest polyline feature using GIS software or programming languages like R or Python. Drawing from this method, Tajima (2003) calculates the distance of each Boston home to the closest highway, finding a positive correlation with property values. In other cases, studies create new polygons that serve as buffers at specified distances from the polyline. In such an instance, Maantay (2007) delineates buffers around highways and truck roads in the Bronx, New York, showing a higher incidence of asthma cases at residential units within the buffer.

3   GPS and CDRs

Urban environments derive their complexity from the movement patterns of individuals, ranging from daily commutes to leisure journeys. In recent years, emerging technologies, such as geolocated logs from GPS and CDRs, have provided powerful tools to analyze these patterns. Figure 2 Panel A provides a visual representation of how GPS data appears when mapped. CDRs—which document telecommunication transactions like phone calls and text messages based on cell tower pings—have become particularly valuable. They offer real-time, detailed insights into human behavior, allowing for comprehensive analyses of dynamic urban environments. These modern tools can be supplemented with machine learning techniques to deliver predictive insights.

Several studies have utilized these data to explore urban mobility. For instance, Ratti et al. (2006) use GPS data from taxis to explore passenger mobility patterns in Singapore. Similarly, Reades et al. (2007) and Kolar et al. (2014) explore how CDR data can be used to map urban dynamics for entire metropolitan regions. GPS data have also facilitated the study of mobility across different sociodemographic groups, helping us understand the patterns of segregation across work and leisure locations (Wang et al., 2018; Xu et al., 2019; Gauvin et al., 2020; Dong et al., 2020; Athey et al., 2021; Moro et al., 2021; Yabe et al., 2023). Figure 2, panel A, illustrates GPS data obtained from mobile devices.

In the case of CDRs, Deville et al. (2014) use anonymized mobile phone data to generate dynamic population maps, providing invaluable insights for urban planning and disaster response. Kreindler and Miyauchi (2021) use CDR data to study commuting patterns and estimate the income distribution within a city. Other work has used CDRs to extract human mobility patterns (González et al., 2008; Song et al., 2010; Kung et al., 2014; Graells-Garrido et al., 2016; Zhao et al., 2022), measure social encounters (Couture et al., 2022), estimate population densities and flows (Deville et al., 2014; Doyle et al., 2014), classify land uses (Soto and Frías-Martínez, 2011), model city structures (Louail et al., 2014), predict socioeconomic indicators (Smith-Clarke et al., 2014), or simulate the spread of diseases (Tizzoni et al., 2014).

The study of walking behavior and local shopping habits in cities has been significantly enhanced by the use of GPS data (Zhu and Levinson, 2015; Vanky et al., 2017). For instance, Hahm et al. (2019) combine GPS data with survey responses to explore how specific features of the built environment in retail areas impact pedestrian activity and shopping habits. In a similar vein, Ponce-Lopez and Ferreira (2021) developed a classification system for commercial spaces based on the appeal they have for individuals and groups, drawing on data from cellphone traces. Such methodologies offer valuable guidance for the redesign and revitalization of retail and commercial zones.

Other studies derive pedestrian trajectories from GPS and pair them with computer vision techniques that measure urban design features. This methodology can help track temporal changes in streets, thereby aiding practitioners in identifying urban blight (Li et al., 2018a; Salazar-Miranda et al., 2021). Abbiasov et al. (2022) use GPS data from 40 million US mobile devices to examine the "15-minute city" concept—whereby all essential services, such as grocery stores and schools, can be accessed within a 15-minute walk from home. This work analyzes how far people travel for amenities and whether having shorter trips, as intended in a 15-minute city, could unintentionally lead to socioeconomic segregation.

GPS and CDR data are offering unique insights into migration patterns. Traditional research methods—such as household surveys, census data, panel data, and qualitative interviews—while valuable, may fall short in capturing the intricacies of dynamic changes in mobility, restricting their ability to provide real-time predictions. GPS and CDR data can yield more nuanced metrics such as travel distances, trip counts, and evacuation distances in real time. Such research is also incorporating metrics rooted in physics. The approach is exemplified by Song et al. (2010), who leverage entropy-based indicators to measure the predictability of patterns in human movement. The development of mathematical and computational models, including Markov chains (Charu et al., 2017) and agent-oriented simulations (Balcan et al., 2009; Heine et al., 2022), have emerged as powerful tools to understand such

phenomena. These advancements allow for a deeper exploration of human mobility dynamics, particularly under external disruptions such as natural catastrophes or pandemics.

Other studies have turned to studying internal migration within cities. For instance, Palchykov et al. (2014) utilize mobile phone data to model human mobility and migration, presenting an innovative approach to studying the social dynamics of cities. Similarly, Blumenstock (2012) used CDR to analyze internal migration patterns in Rwanda. Another study by Li et al. (2023) uses similar data to track people's movements in Sierra Leone, demonstrating how different groups might be exposed to COVID-19. The findings show that people with lower incomes traveled more during the pandemic, potentially increasing their risk of catching and spreading the virus.
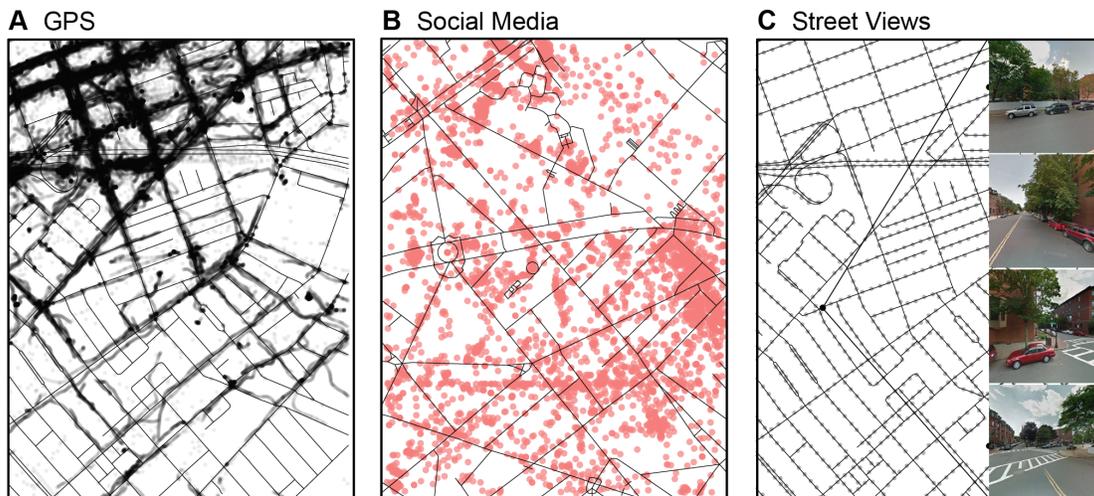
FIGURE 2: Panel A shows GPS data collected from smartphones used in Salazar-Miranda et al. (2021). Panel B shows social media data from Twitter used in Salazar-Miranda et al. (2022a). Panel C shows street view data obtained from Google.

4   LARGE TEXTUAL REPOSITORIES

Large textual repositories—encompassing sources such as internet posts, newspapers, online forums, advertisements, and public records—serve as a rich data resource to study a myriad of social and urban phenomena (Gentzkow et al., 2019). When combined with advanced text mining techniques and natural language processing (NLP), these data can reveal patterns, trends, and sentiments that traditional sources might miss. Using text mining techniques, researchers can gain powerful tools to monitor market trends, forecast behavior, and ultimately,

shed light on the socioeconomic dynamics that shape our cities.

Sentiment analysis, one of the applications of text mining, can quantify the public's feelings toward specific neighborhoods or property types (Adams and McKenzie, 2012; Jenkins et al., 2016; Lai and Kontokosta, 2019; Yang et al., 2022). For instance, Hu et al. (2019) leverage this technique to analyze neighborhood online reviews. Similarly, sentiment analysis has been used to gauge public attitudes toward sustainable practices such as green buildings or retrofitting initiatives. Shen and Li (2023) use text mining to examine public sentiment regarding green housing in China, finding that positive sentiments are motivated by ecological, environmental, social, and individual benefits. In contrast, negative sentiments are primarily associated with price and quality issues. Text mining has also been applied to explore the indoor environmental quality (IEQ) of specific properties; for example, Villeneuve and O'Brien (2020) use this technique to study more than one million Canadian Airbnb reviews, gleaning new insights about IEQ complaints, their seasonal trends, and the consequent impact on guest satisfaction.

Extracting urban proxies from textual data presents another approach to understanding local characteristics. Saiz and Simonsohn (2013) show that document frequencies in large decentralized textual databases—such as the internet and newspaper repositories—can capture the cross-sectional variation in the occurrence frequencies of social phenomena. These authors characterize the statistical conditions under which such proxying is likely. They successfully use this approach to proxy for a number of major economic and demographic variables across US cities. They also obtain document frequency measures of corruption by US state and the country and replicate the econometric results of previous research studying the covariates of past corruption indexes. They subsequently provide the first measure of corruption in American cities, demonstrating that poverty, population size, service sector orientation, and ethnic fragmentation predict higher levels of corruption in urban America.

By analyzing the relative frequency and context of specific words or phrases across space, researchers have created proxies that reflect a variety of factors, from the quality of green spaces to economic activity (Glaeser, 2018; Bendeck and Andris, 2022) and accessibility to public transport (Liu et al., 2017; Markou et al., 2019). For example, Saraiva et al. (2022) use text analysis of official police data to identify spatial crime patterns in Porto, Portugal. Similarly, Ghahramani et al. (2021) analyze reviews from TripAdvisor to assess the quality of urban green spaces.

Parallel literature has used search volume indexes (SVIs) as indicators of local trends. For instance, spikes in search queries related to flu remedies in a city can indicate an increase in local infection rates (Ginsberg et al., 2009; Polgreen et al., 2008). Similar results have been

obtained regarding searches about COVID-19 (Lu and Reis, 2021). Going beyond health concerns, Choi and Varian (2012) argue for a broader utilization of SVIs to contemporaneously estimate a variety of social, economic, and epidemiological phenomena across urban areas. Meanwhile, Saiz and Simonsohn (2013) contrast proxies derived from internet word frequency against those based on SVIs. It is important to note that the former reflects the supply of information available in large textual repositories, while the latter is based on the public's demand for information via search engine queries. Their research suggests that demand-driven measures work better in contexts where searches correspond to individuals seeking localized goods or services that cater specifically to the searched community. However, they also find that SVIs are less reliable when the motivation behind a search doesn't directly relate to the consumption of a good or service distinctly associated with the queried term.

Large textual repositories not only offer insights into neighborhood attributes, but they also can shed light on a city's regulatory environment. Stacy et al. (2023) applies machine learning algorithms to assess US newspaper articles, exploring the effects of land use reforms on housing density. The results show that eased regulations amplified high-end housing development. Conversely, stricter regulations pushed up median rents and curtailed affordable housing availability. In a complementary study, Mleczko and Desmond (2023) employ NLP to public records to measure exclusionary zoning across the US.

5   Data from Social Media

Increasingly, researchers are turning to social media data—both textual and non-textual—to uncover patterns of network formation and explore urban economic or social behaviors. Some studies extract direct information from posts on social media, using similar approaches to those described in the previous section. For instance, Mitchell et al. (2013) developed a geo-tagged dataset for over 80 million Twitter keywords, using it to categorize cities based on word use similarities. They also construct a happiness index for cities, correlating it with local demographics. Figure 2 Panel B shows data derived from social media platforms, such as Twitter. Similarly, Haranko et al. (2018) analyze advertised skills on LinkedIn, comparing male and female self-reported attributes to citywide gender gaps.

Social network data can also shed light on geographical network densities. For instance, Takhteyev et al. (2012) examine global Twitter data, revealing that most social connections tend to remain within the same metropolitan area. They find that factors like physical distance, frequency of airline flights, national borders, and linguistic differences play pivotal roles in shaping inter-city Twitter ties. Delving into Facebook's "friend" networks across US

colleges, Traud et al. (2012) discover that shared high school experiences greatly influence connections, even more than shared classes, residences, genders, or other attributes.

Laniado et al. (2018) analyze Tuenti—a Spanish social network with around 10 million active users—finding that spatial distance hampers the likelihood of forming friendships, especially among younger users. However, once these distant friendships are formed, the intensity of the interactions remains constant regardless of distance. Interestingly, friendships within densely connected groups tend to arise over shorter spatial distances than do connections between individuals from distinct groups.

Obradovich et al. (2022) use Facebook's marketing application programming interface (API) to download information about self-reported interests, clicking behavior, "likes," software downloads, GPS location, and activities on other websites featuring Facebook ads. Their comprehensive data extraction, covering two billion Facebook users from 225 countries and territories, resulted in roughly 60,000 topic dimensions. This vast pool of information offers insights into the cultural similarities and differences across regions.

Additional work studies the geographic nuances of user behavior *inside* the platforms. For instance, Agrawal et al. (2015) explore a crowdfunding platform that connects artists with funders. Their findings show that artists tend to receive larger investments from local funders with whom they share offline social ties. However, the influence of these localized social ties does not persist past the first investment, implying that investment behavior converges to a market-like equilibrium across cities, beyond the scope of these conventional social networks.

Researchers also increasingly use social media to anticipate offline behavior across cities. As an example, DiGrazia et al. (2013) show that the intensity of Twitter messages can predict election outcomes in US congressional districts. In the context of the COVID-19 pandemic, Kuchler et al. (2022) use Facebook data, revealing a stronger likelihood of the virus spreading between regions that exhibited stronger social network connections. This suggests that regions with stronger online social network connections are more likely to experience faster virus spread, implying that offline personal interactions often align with online ties. In terms of human mobility, Hawelka et al. (2014) analyze over a billion geolocated tweets from 2012 to estimate international traveler volumes based on country of origin. Their findings, when juxtaposed with additional data sources, highlight the potential of Twitter as a tool for quantifying global mobility patterns. Furthermore, Twitter has been instrumental in exploring intra-city mobility, with recent work examining segregation patterns and street-level social activity (Heine et al., 2021; Salazar-Miranda et al., 2022a).

6   Credit Card Data

Credit card data can provide real-time, granular insights into economic activities within cities. Each transaction serves as an individual snapshot of consumption habits, forming a high-resolution map of urban economic dynamics. Unlike periodic surveys or censuses, credit card data provides a continuous stream of information, offering more nuanced perspectives on urban life. Furthermore, the inherent geographical specificity of credit card data enables precise spatial analysis, which is critical for urban planning and real estate analysis. Consumption patterns linked to geographic locations can also shed light on socioeconomic disparities within a city. The kinds of products people buy, the locations of their purchases, and even the timing of their transactions can all provide insights into social heterogeneity. For instance, researchers have utilized geolocated credit card records to study how individual mobility patterns are influenced by factors such as gender, age, and social habits (Lenormand et al., 2015; Di Clemente et al., 2018) and have explored their predictability and regularity over time (Krumme et al., 2013).

Other studies focus on predicting shopping patterns by identifying popular retail locations (Sobolevsky et al., 2014), times of peak activity (Singh et al., 2015), and the stores that are frequently visited together (Yoshimura et al., 2018). The findings can help city authorities manage urban areas more effectively and enable retailers to improve their services based on their customers shopping habits. For instance, Sobolevsky et al. (2014) use a dataset of over 10 million anonymized bank card transactions from Spain to explore human mobility and consumption patterns. They find that the transaction data can be used to delineate regions of economic activity, with apparent differences between the spending patterns of residents and foreign visitors.

Credit card data can also provide granular insights in real time to study changes in consumer spending across different categories (Bounie et al., 2023) and how changes in gasoline prices affect consumers' mobility and expenditure patterns (Carvalho, Vasco M. and Hansen et al., 2020).

Municipalities around the world are keen to understand the impact of retail on local taxation, job creation, and traffic patterns, and credit card data is increasingly becoming a valuable tool for such analyses. Farrell et al. (2017) examine offline credit card and debit card transactions in Detroit and New York. Their research identifies two critical insights. First, Detroit and New York residents made 71.6 percent and 56.8 percent of their purchases, respectively, outside their 20-minute neighborhood. Second, lower-income families travel further to shop. Nevertheless, between 2013 and 2016, the distance between residents and

stores diminished, especially for low-income residents.

Another important question concerns the continued rise of online shopping and its potential to overshadow local retailers. Drawing from comprehensive data on daily credit card activities, Relihan (2022) observes that consumers allocate the saved time by online shopping to other more time-consuming activities, like visiting cafes. This suggests that, paradoxically, certain physical retail outlets might be reaping benefits from the surge in e-commerce.

7   Street View Imagery and Online Pictures

There has been a surge in research focused on the impacts of urban design, visual attributes, and the physical appearance of cities, enabled by vast repositories of street imagery. Recent work has cataloged the numerous such studies across a wide array of disciplines, from vegetation and transportation analysis to health and socioeconomic research (Biljecki and Ito, 2021; Zhang et al., 2023).

Since its launch in 2007, Google Street View has become the primary resource for street-view imagery due to its extensive database of georeferenced pictures. Encompassing hundreds of thousands of miles, this resource offers an unparalleled look at streets and the socioeconomic forces that shape them. Figure 2 shows an example of mapped street-view imagery. It has been applied in diverse domains such as physical environment audits (Kang et al., 2018; Li et al., 2018b; Zhang et al., 2020; Ning et al., 2021), urban mobility and transportation (Lu et al., 2019; Hong et al., 2020; Mooney et al., 2020), energy use estimation (Liu et al., 2019; Zhang et al., 2022; Sun et al., 2022), and real estate (Yang et al., 2020; Johnson et al., 2020). There is also a growing trend of using crowdsourced street-view images, facilitated by platforms like Mapillary and KartaView.

Street-view imagery can offer a wealth of information for analyzing urban aesthetics and architecture. These images capture intricate details of buildings, materials, architectural styles, and other urban features such as green spaces, abandoned structures, and public art (Less et al., 2015; Zou and Wang, 2021; Zhang et al., 2018). The study of these elements provides insights into amenity values, which play a significant role in determining property prices (Law et al., 2019; Kang et al., 2020; Lindenthal and Johnson, 2020; Qiu et al., 2022). Furthermore, this imagery offers a platform to evaluate the impact of architectural changes over time, offering vital data for urban planners and real estate developers. An exemplar of this potential is the *Treepedia* project,[1] where a deep learning model, trained on Google Street View images, quantifies the tree canopy of streets across 30 cities, providing a scalable

---

[1]http://senseable.mit.edu/treepedia

alternative to manual audits (Seiferling et al., 2017; Li and Ratti, 2018).

Moving beyond the visible, street imagery also provides insights into latent aspects, such as crime rates (Khosla et al., 2014), or shifts in human dynamics (Zhang et al., 2019). For example, Naik et al. (2017) link changes in urban appearance with economic and demographic data across five US cities, delving into factors driving neighborhood improvements.

Complementing the street-level perspective, point-to-point online images from platforms like Flickr and Instagram provide unique, personal perspectives of urban environments (Li et al., 2013). These user-contributed images are continually updated, geographically diverse, and often capture perspectives that might be absent in datasets that are collected more systematically. Such images offer granular details about urban beauty, architectural style, and individual experiences within urban spaces. By analyzing these photos, researchers have identified picturesque areas (Seresinhe et al., 2017a,b), distinguished architectural styles (Saiz et al., 2018), assessed the appeal of cities (Paldino et al., 2016), and gauged public sentiment toward different urban environments (Quercia et al., 2014; Feick and Robertson, 2015; Preis et al., 2020). Notably, Saiz et al. (2018) use this rich online photo archive to gauge how people value the aesthetic elements of their environment, establishing a scalable metric for attractiveness both across and within cities.

8   Sensors

Scientists and policymakers have adopted street sensors to measure specific aspects of the urban environment (Ang and Seng, 2016). These devices produce vast amounts of data that can be used for research, policy, and law enforcement. Users typically have information regarding the geolocation of the sensor (which could be stationary or mobile), the timestamp of the measurement, and the measurement itself.

Well-established temperature, air pressure, and precipitation measurement stations serve as an essential resource for urban environment monitoring in cities around the world. The study by Muller et al. (2013) reviews the literature on the use of these stations and documents the outstanding challenges. These include the need for enhanced documentation of sensor networks, standardized methodologies, financial barriers, and the establishment of the long-term global networks essential for advancing urban climate research.

Pollution monitoring stations constitute another critical sensory data source. A notable example is the air pollution sensors established in numerous American embassies around the world, which broadcast real-time data via Twitter. In countries with limited public transparency, this data source has been instrumental in increasing citizen advocacy, subsequently

enhancing air quality (Jha and Nauze, 2022).

Noise-measuring sensors also play an important role. Alsina-Pagès et al. (2021) use sensors in Girona, Spain, to quantify the auditory impact of COVID-19 lockdowns on the city. The largest noise reduction was experienced in central areas during the evenings and nights, attributable to the temporary closure of restaurants and bars.

Motion sensors—which can count pedestrians (Schauer et al., 2014) or vehicles—function similarly to GPS and CDRs but operate on a more localized scale. A creative application by Kishino et al. (2018) involved installing motion sensors in garbage collection trucks in Fujisawa City, Japan, to estimate waste volumes.

Light sensors also hold promise for future urban research. Presently, they enhance smart city technologies by modulating public lighting based on current visibility conditions (Sikder et al., 2018).

Given the ongoing reduction in sensor costs and the proliferation of Internet-of-Things devices—which also capture sensory data—we anticipate that such data sources will become more prevalent in urban research.

9 Volumetric Data and Street Shapes

Two fundamental aspects of urban morphologies—building volumes and street shapes—are crucial to urban planning, design, and real estate analysis. Volumetric data, which represents the three-dimensional structure of urban spaces, include attributes such as building heights, density, and massing (Berghauser Pont, M.Y.; Haupt, 2009). Figure 3 Panel A provides an illustrative example of LiDAR data. These data offer insights into the spatial distribution of built volumes in cities, supporting analyses of urban form, skylines, solar access, and ventilation.

Research into the volumetric characteristics of built environments is gaining popularity. Some studies use high-resolution airborne LiDAR data to distinguish features like building massing, height, volume, exposed surface area, and roof geometry (Bonczak and Kontokosta, 2019; Sánchez-Aparicio et al., 2020). Additionally, the morphological properties of informal settlements have been examined using terrestrial LiDAR data. By implementing an automated and scalable approach in Rio de Janeiro's largest favela (informal settlement), variations and commonalities in street morphology were identified. These findings contribute to the creation of detailed maps that can guide urban planning decisions related to crowding, safety, air quality, and accessibility (Salazar-Miranda et al., 2022b).

Hedonic models in real estate are increasingly utilizing volumetric data on specific build-

ings and ensembles. Yu et al. (2007) use 3D models of buildings in Singapore to determine the presence of sea views in private high-rise residential apartments, showing that they contribute an average premium of 15 percent to the property price. Lindenthal (2020) uses the variety and contrast of building volumes from the city of Rotterdam's 3D building model. The findings support the hypothesis that architectural homogeneity is positively valued in the residential property market.

On another front, street shapes create the network of paths that connect individuals and neighborhoods. They play a central role in urban mobility, accessibility, land use patterns, and social dynamics (Hillier and Hanson, 1984; Frank and Pivo, 1994; Cervero and Kockelman, 1997; Hillier and Hanson, 2005; Marshall, 2005). In their seminal work, Hillier and Hanson introduce the "space syntax" concept, emphasizing the influence of spatial configurations on pedestrian and vehicular flow patterns (Hillier and Hanson, 1984). Marshall (2005) builds on this foundation, examining the impact of different street formats on urban vitality.

More contemporary research has tapped into the latest data to delve into street characteristics (Boeing, 2021). Using road network data, Boeing (2020), Louf and Barthelemy (2014), and Barrington-Leigh and Millard-Ball (2015) categorize and measure street network layouts and orientations across various cities. This work illustrates the distinct spatial qualities of cities, spotlighting the role of history, geography, and culture in shaping urban designs.

Researchers have also connected street configurations to outcomes in urban mobility and social dynamics. Studies show that grid-patterned, interconnected street networks enhance walkability and decrease vehicular ownership rates (Ewing and Cervero, 2010; Barrington-Leigh and Millard-Ball, 2017). Grids are also shown to facilitate land subdivision and promote denser development patterns (Conzen, 1960; Siksna, 1998; O'grady, 2014). Additional studies indicate that the form of street networks correlates with factors such as vehicle ownership and single-occupancy vehicle travel (Cervero and Kockelman, 1997; Barrington-Leigh and Millard-Ball, 2017); walkability (Hajrasouliha and Yin, 2015; Salazar-Miranda et al., 2021); community sentiment (Brown and Cropper, 2001; Nasar and Julian, 1995; Rogers and Sukolratanametee, 2009; Wood et al., 2010); social integration (Grannis, 1998; Salazar-Miranda, 2020); and social capital and the formation of ties (Cabrera and Najarian, 2015; Leyden, 2003; Ziersch et al., 2005; Farrell et al., 2004; Hipp and Perrin, 2009). Furthermore, the design nuances of street layouts—such as cul-de-sacs versus elongated streets—can influence housing prices (Asabere, 1990; Song and Knaap, 2003). More recently, Salazar-Miranda (2022b) developed a measure for neighborhood design based on street network data across US neighborhoods. This measure is used to evaluate the impact of design on environmental and social sustainability, showing that certain designs increase greenhouse gas emissions, amplify

social detachment, and promote sedentary behaviors.

10 PUBLIC TRANSPORTATION AND OTHER VEHICULAR DATA

Transportation systems are the guts of the metropolis. After all, cities arise precisely in order to facilitate the transportation of goods, people, and ideas (Hall, 1998). Strides have been made in assessing transportation activities within cities, particularly concerning worker commutes.

Historically, most studies relied on surveys detailing the origins and destinations of commuters, modeling home, job, and modal choices—either individually or jointly—using discrete choice models. For example, Manaugh et al. (2010) examine the Montreal Origin-Destination Survey, which encompassed over 30,000 home-to-work automobile trips. They determined that unseen heterogeneities concurrently impact both the choice of home-work location and commuting distances. Time use surveys offer another fruitful data source for understanding commuting and leisure trips globally, as well as their interactions with labor markets and urban layouts (Charmes, 2015). Unfortunately, in many countries, these samples lack the granularity to deliver significant data on city or metropolitan scales.

More recent work is incorporating new data sources. As highlighted earlier, CDR, sensor, and GPS data can inform studies on individual commuting behaviors (Papinski et al., 2009). The advent of electric and autonomous vehicles is significantly amplifying the volume of GPS data collected from commutes. While much of this data is proprietary or confidential, some open-source repositories are accessible to researchers (Yin and Berger, 2017). However, these datasets are not yet expansive enough to comprehensively analyze commuting patterns. This remains an area of promising growth.

Turnstile data, which record the time and location of subway system entries, have also been used to evaluate the localized and time-specific demand for public transportation (Reddy et al., 2009). Similarly, data logs from taxi and ride-sharing services like Uber, Lyft, and Didi, are producing insights into urban mobility (Henao and Marshall, 2019; Barajas and Brown, 2021). However, these companies typically restrict data access.

The literature on urban transportation is vast (De Witte et al., 2013; Chatterjee et al., 2020), and we cannot do it justice in this piece. However, a novel and easily accessible data source has emerged that can be extremely useful to social scientists. Instead of focusing on individual trips and choices, researchers may simply be interested in aggregate commuting statistics. Both Akbar and Duranton (2017) and Akbar et al. (2022, 2023) propose using real-time data from the Google Maps API. This source delivers aggregated CDR data on traffic

conditions, encompassing general trends and specific origin-to-destination routes within cities. In China, similar insights can be gathered through partnerships with Baidu (Gu et al., 2021). Utilizing Google Maps, Saiz and Wang (2023) extract driving times for a large representative sample of origin and destination points in the Boston metro area. By overlaying these data with the US Census travel survey's dyadic commuting estimates, they document significant impacts of geography on local commuting outcomes. Their estimate indicates that 13.73 percent of central city traffic results from trips that have to bypass the ocean through the urban core.

11   Historical Data

There is a growing interest in recycling historical data to furnish large quantities of historical information with localized precision. One example is the digitization of ancient maps into GIS formats. Similarly, large quantities of historical records and documents, many of which were initially handwritten, are being scanned and subsequently transformed into digital formats via machine learning techniques (Shen et al., 2021). When these records house addresses, they can be combined with maps to pinpoint the exact geolocations of families, businesses, and POIs. Figure 3 Panel B demonstrates the georeferencing of historical maps. Moreover, records that mention individuals with matching names and characteristics, which presumably denote the same person, are also increasingly matched across time periods (Abramitzky et al., 2021). By geocoding linked individual data from varying years, there's an opportunity to decode early migration trends.

Shertzer et al. (2016) digitized the first zoning map of Chicago from 1923, finding that old African American neighborhoods were subsequently zoned at higher densities than other areas. Following a similar vein, Twinam (2018) digitized a series of zoning maps of Seattle spanning from 1920 to 2015, allowing for a comparative study of outcomes across places that were subjected to zoning changes.

Furthering this trend, Salazar-Miranda (2022a) digitized urban master plans from the United States Housing Corporation (USHC)—the first housing initiative funded and promoted by the US federal government, from 1918 to 1919. The USHC planned 55 neighborhoods, establishing rules that dictated the street layout, public space allocation, and building design and placement. However, only 22 of the 55 USHC neighborhoods were executed, allowing for a comparative analysis between the built and unbuilt projects. Salazar-Miranda (2022a) finds that USHC neighborhoods have changed in some dimensions, but the original street layout has endured as in the original plans.

Tapping into another rich source of historical data, Sant'Anna and Cortes (2023) use geolocated microdata from the US censuses of 1930 and 1940 to identify houses that appeared in both years. Their findings show that the Mexican Repatriation had a negative impact on housing prices in predominantly Mexican American neighborhoods. Ciccone (2021) utilizes the official casualty lists from World War I in Germany to, which include the first and last names of the deceased and their municipality and date of birth, to show that early local population losses were persistent through the 20th century. War records are also used by Costa et al. (2018), who matched them to historical US Census data to show that people who fought together during the American Civil War (1861-1865) co-located in the same towns.

Back in Europe, Chambers et al. (2021) constructed a dataset based on handwritten records of the property holdings and transactions from the early 20th century for King's College and Trinity College in Cambridge, England, and Christ Church and New College in Oxford, England. They find the long-term real returns to residential real estate (1901-1983) to be low, at 2.5 percent. Meanwhile, Drelichman et al. (2021) analyzed 67,521 individual records from the Spanish Inquisition, between 1504 and 1700. By geocoding various details like residence, birthplace, and location of alleged offenses, they managed to map 57,924 trial observations to contemporary municipalities in Spain, subsequently finding negative effects of inquisitorial activity on today's local social and economic outcomes.

The potential for the use of geocoded historical data in the urban sciences is immense, and their creation, wider dissemination, and use are exploding.

12   Governmental Data

Arguably, the most prevalent and least novel governmental data sources have received extensive attention in the literature, as showcased in works such as Barkham et al. (2022), Hielkema and Hongisto (2013), Vetrò et al. (2016), Connelly et al. (2016), and Wilson and Cong (2021). Consequently, our discussion on this topic is brief.

Census data and national or regional surveys are important data sources in most countries. In certain nations, such data are reported at sub-municipal levels, such as ZIP codes or census districts. Depending on the specific country, these data might encompass details on population, age distribution, ethnicity, income levels, local housing markets (e.g., housing rents), and many other relevant demographic and economic characteristics.

Other urban-centric data can be obtained from sources such as public registries, birth and death records, legal and land registration records, the postal service, open data platforms maintained by municipalities, city budgets, tax records, archival records, and more. Evidence

A Volumetric **B** Historical **C** Volunteered
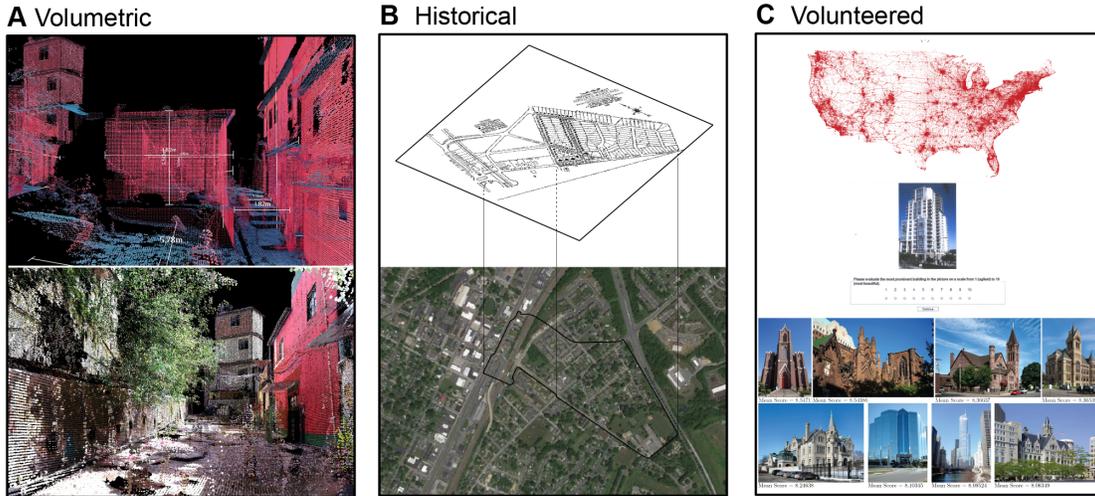
FIGURE 3: Panel A shows LiDAR data from Salazar-Miranda et al. (2022b). Panel B shows historical maps digitized by Salazar-Miranda (2022a). Panel C shows both the density of volunteered image data from Flickr on a map (top) and an image rating survey (mid and botttom) as used in Saiz et al. (2018).

of a global trend towards increased digitization and accessibility of records, facilitated by optical character recognition technologies, is clear (Malik et al., 2022).

A few countries are at the forefront, offering microdata to researchers under confidentiality agreements. Notably, Scandinavian countries collect an enormous volume of data on their entire populations due to several factors: i) their extensive public service offerings (encompassing education, health and medical services, higher education, Social Security, unemployment benefits, maternity support, and more); ii) their robust taxation systems essential to sustain such services, which entail the collection of income, business, and labor-related data; and iii) the adoption of a national ID, facilitating the linkage of different records.

For example, Edin et al. (2003) and Damm (2009) investigate the impact of ethnic enclaves on immigrant labor outcomes using microdata from Sweden and Denmark, respectively. These data identify the exact residence of immigrants and labor metrics such as wages and employment. Both studies leverage policies that randomize the location of new asylum seekers, finding improved labor market outcomes for immigrants who were assigned to ethnic enclaves—presumably because they received help from their co-ethnic networks to find jobs.

Although the degree of data centralization seen in the likes of Sweden and Denmark remains rare, many countries are making advancements in interlinking and geocoding their available datasets. For instance, the US has made progress in linking longitudinal data from

various sources, including the Census, Social Security, and state unemployment records. Such integrations have paved the way for services such as IPUMS, which allows for public use of microdata. Additionally, advancements in historical censuses, such as the Census Linking Project, are providing new avenues for understanding demographic shifts and socioeconomic patterns. Such data have allowed for the emergence of new literature that examines differences in upward mobility by neighborhood and its potential constraints (Chetty et al., 2014, 2018)

13   BUSINESS AND MARKETING

Private firms maintain a vast repository of data stemming from their accounting, investment strategies, operations, sales, and records of customer online behaviors. While these data are predominantly proprietary and not readily disclosed, their use in academic research has increased exponentially, especially within economics and finance (Einav and Levin, 2014). Nevertheless, the application of such datasets by urban scientists remains in its nascent stages.

Sometimes, researchers can access open data with urban applications from publicly listed companies. For instance, Dougal et al. (2015) identify the headquarters of all publicly listed corporations in the US and show that their corporate investment decisions are affected by local trends. In another study, Fairweather et al. (2023) work with data from an experimental initiative by Redfin, an online real estate firm. A group of potential customers browsing the website were presented with flooding risk details, which reduced the probability of home hunters to bidding for the most affected properties.

Public-sector surveys also offer industry-specific data. For instance, Li and Li (2013) use the Annual Survey of Industrial Firms by China's National Bureau of Statistics, spanning 1998-2007. Coupled with data on newly constructed major highways, they show that establishments in cities with increased highway investments decreased their output inventories in storage, capitalizing on the ease of just-in-time deliveries.

Online commerce platforms also prove to be fertile grounds for data collection. Hortaçsu et al. (2009) collected online data from the US-based Ebay and collaborated with MercadoLibre, a Latin American equivalent, to obtain proprietary transactional records. Their findings still show a negative impact of geographic distance between buyers and sellers on online trading, though to a lesser extent than in traditional offline commerce. Meanwhile, Mast (2023) uses data from the marketing firm Infotutor, which collects household details by centralizing mailing lists from diverse private entities, such as magazine subscriptions. This paper finds that new housing construction ultimately increases the availability of medium- and low-

income apartments via a chain of relocations and subsequent vacancies. These results have been replicated by Bratu et al. (2021) using data from Finland and are consistent with the broader literature in housing affordability that claims positive effects of housing construction, recently summarized in Saiz (2023).

In sum, the integration of corporate data into urban research presents a promising frontier. In situations where firms' reservations can be minimized, confidentiality maintained, and ethical standards upheld, we see this as a growing opportunity for researchers.

14   REAL ESTATE TRANSACTIONS DATA

A major research literature in economics and finance, initiated by Rosen (1974) uses hedonic models to value specific attributes of residential and commercial properties. Real estate prices not only interest economists but also capture the private valuations of critical urban and environmental aspects (Malpezzi et al., 2003). While we have already discussed how property data can be integrated into other sources, we analyze its origins in more detail below. The key feature of these data is that property addresses can be geocoded, pinpointing the precise latitude and longitude of each property (Zandbergen, 2008).

14.1   Public Registries

Some national or regional statistical offices collect city-level average housing price/rent data. On a micro transaction level, regions that duly document property sales and deem such information public often disclose registry records. In the US, several private companies consolidate data records from local jurisdictions. While there is a fee for accessing these processed public data, their availability has spurred numerous studies on urban housing markets (Hill, 2013). Zillow, for example, uses these data to publish housing price indexes at the ZIP code level, as seen in studies like Ramani and Bloom (2021), who estimate relative declines in the prices of housing in central and denser areas during the COVID-19 pandemic. Some countries like Singapore also provide public registry data access, facilitating research such as Wong (2014), who identifies preferences for ethnic segregation in the city's housing market.

In some countries, property transaction data are available, but the true prices are often underreported to evade taxation (Anagol et al., 2022). This leads to a disparity between the officially recorded values and the actual contract prices utilized for monetary exchanges between sellers and buyers (Shimizu et al., 2012).

24

14.2 Brokers' Records

Brokers maintain records on contract prices from their past sales or rentals, providing another source of information. An example is Bracke (2015), who uses the property agent's records to compare the price-to-rent ratios evolution across central London. Similarly, Wong (2008) obtained data from the Centaline Property Agency—a major local broker in Hong Kong—to document a price reduction in condo apartments in areas severely impacted by the 2002-2004 SARS outbreak.

In certain regions, broker coalitions form to share such data. For instance, Lazrak et al. (2014) use data from the Dutch Association of Real Estate Agents to show that buyers pay an additional 26.9 percent for homes that are listed on the Netherlands' historic heritage list.

In the commercial real estate arena—covering office, retail, and industrial properties— large consolidated brokerage and services' firms have historically dominated data collection for rentals and pricing. Kok and Jennen (2012) integrate the proprietary transaction databases of the largest real estate agents in the Netherlands: CBRE, DTZ Zadelhoff, and Jones Lang LaSalle. By merging these databases with information on each building's energy features, they show that buildings designated as inefficient (rated D or worse on the EU energy performance certificate) command rental levels that are 6.5 percent lower than their energy-efficient counterparts (rated A, B, or C).

14.3 Asking Prices from Advertisements

Real estate ads are another data source for researchers. Although these ads can be mined to offer insights about housing prices and rents, the data might be somewhat inconsistent over periods and between different ads. Moreover, these ads usually reflect the initial asking price of a property, not the final contract price.

Nonetheless, some of these errors may cancel out with regard to changes across long time periods. For example, Garcia-López et al. (2020) analyze data from Barcelona's largest housing advertisement portal, Idealista. Their findings show that neighborhoods with high shares of Airbnb short-term rentals experienced faster housing price and rental appreciation between 2012 and 2017. Using the same web-scraping approach on the same source, Kholodilin et al. (2022) find that introducing rent control in Catalonia led to a reduction in rents. However, monthly dwelling starts declined by 6 percent during the rent control period, while other Spanish regions experienced an almost 12 percent increase.

14.4 Multiple Listing Services

In certain areas, associations of property agents maintain centralized records for all advertised properties, whether for sale or rent. These multiple listing services (MLSs) consolidate asking price data and offer platforms that streamline the collection of other property information, including changes in asking prices, time to market, detailed property characteristics, broker names and affiliations, and any advertisement write-ups. While primarily designed to facilitate information flow and transactions between brokers, researchers often use these platforms. For instance, Schaefer (1990) uses the Ontario MLS to find that designating specific properties as belonging to a floodplain area does not seem to have a clear effect on their valuations.

14.5 Appraisals

Professional assessments, which estimate specific buildings' market value, are commonly conducted by mortgage lenders, local governments, investors, and property buyers. If the comparison properties—those recently traded—are reliable, this method can provide accurate approximations. However, appraisals can be biased—or at least conditionally biased—in both the residential (Cho and Megbolugbe, 1996; Calem et al., 2021) and commercial arenas (Geltner et al., 2003; Geltner, 1989). Therefore, the potential biases must be weighed against the large quantities of data "signal" they provide (Clapp and Giaccotto, 1992). In general, appraisals accurately capture long-term relative growth (Bourassa et al., 2006). Moreover, statistical techniques can be used to rectify their short-term biases (Geltner, 1996; Fisher and Geltner, 2000).

For example, Rico-Juan and de La Paz (2021) use professional appraisals in Spain, as validated by McGreal and Taltavull de La Paz (2012), to predict housing values via a random forest model. This model outperforms conventional log-linear hedonic specifications.

14.6 Mortgage Lenders

Mortgage lenders routinely process and store data on housing and commercial real estate, potentially facilitating scientific research. In some countries, lenders are mandated to publicly disclose all pertinent information about securitized mortgages. Researchers occasionally use these data to infer the evolution of urban outcomes. Schuetz et al. (2008) demonstrate that proximity to properties in foreclosure is associated with decreasing neighborhood valuations. Of course, there is large and separate literature in finance using these data to study mortgage default (Foote and Willen, 2018) and pricing (Boyarchenko et al., 2019), although not all

studies give spatial issues the attention they deserve. However, some investigations, such as those focusing on redlining or lenders' challenges in catering to low-income neighborhoods, do delve into these aspects (Lang and Nakamura, 1993; Aaronson et al., 2021; Conzelmann, Claire; Salazar-Miranda, Arianna; Phan, Toan;Hoffman, 2023).

14.7   Data Aggregators

In some cases, professional data aggregators obtain information from all the aforementioned sources to deliver consistent datasets on rent and prices. These aggregators may also tap into local data streams, such as contact with brokers, local trade publications, or surveys. In the commercial arena, the Costar Company has arisen as a leading data aggregator in both the US and the UK. Fuerst and McAllister (2011) use this source to discover that, compared to other buildings in the same sub-markets, eco-certified buildings display both a rental and a sale price premium.

15   CROWDSOURCED AND VOLUNTEERED DATA

Throughout the text, we've mentioned several crowdsourced datasets, such as user-uploaded images and social media. Nonetheless, we believe it is relevant to single out urban crowdsourced data as a distinct and important resource to urban researchers and practitioners. Crowdsourced data are not centralized by a single private or public data provider but arise from the decentralized actions or behaviors of thousands or millions of users. Niu and Silva (2020) identify 226 relevant papers that employ crowdsourcing in the quantitative urban literature.

We can categorize crowdsourced data into two main types: volunteered and behavioral. Volunteered crowdsourced data is explicitly solicited, where the public is asked to contribute information (Goodchild, 2007). This data collection process is rooted in intentional online community-based involvement, which might either be long term or ad hoc.

A salient example of a long-term volunteered source is OpenStreetMaps. It encompasses billions of data points, detailing roads, POIs, and natural geographic features around the world. This dataset can be viewed as the topographic and urbanistic equivalent of Wikipedia. While volunteered geographic information might have imperfections, it is generally accurate for most statistical analyses in the social sciences and urbanism (Haklay, 2010; Ciepłuch et al., 2010; Fan et al., 2014). However, the quality of contributions can vary among users, as described by Arsanjani et al. (2013). In terms of policy, platforms like CitySourced allow citizens to report problems with streets, urban infrastructure, and noise. Municipalities can

then use these data to identify and rectify hotspots.

Volunteered data can also arise from specific researcher-driven efforts on a particular topic. One example is in the gathering of information on informal transit systems and bus routes, where data are usually lacking (Williams et al., 2015; Yun et al., 2019). Another method involves data hackathons, where participants contribute data or software over short periods, like one or several days. In such a hackathon, Masdeval and Veloso (2015) add an urgency feature to the platform SeeClickFix, a tool that lets citizens report urban problems, and then analyze the results. Often, these specific data collection methods require compensating the contributors, either through survey platforms like Qualtrics or online job platforms like Amazon Mechanical Turk. An example is Saiz et al. (2018) using the former to assess the visual appeal of buildings, for which participants were compensated. Figure 3 visually depicts the images and survey used in the analysis.

16   Challenges of Urban Big Data for Researchers

Research using the types of data that we have described in this piece may suffer from the potential pitfalls of any scientific endeavor: malicious practice, subjectivity, ideologization, irreproducibility, lack of relevance, poor in-sample fit, poor out-of-sample predictions, lack of external validity, lack of clarity, lack of causal attribution, and poor modeling choices, among others. In addition to these generic challenges, we believe that urban big data may be more prone to other specific problems related to data collection, analysis, and inference. In this section, we detail some of them.

Awareness is the first step to minimize challenges. In that spirit, we provide only a brief description of the problems and a few suggestions of sources for further study. We advise researchers to delve deeper into these issues. They should also perform their due diligence to identify and address them. Below, we find it productive to sometimes frame research issues in the context of a "treatment and control" framework, where we are trying to ascertain the impact of a "policy" variable on an outcome (Imbens and Rubin, 2010).

16.1   Modifiable Areal Unit Problem (MAUP)

Many urban data sources provide geolocated observations. To conduct analyses, researchers often need to define geographic boundaries to contrast "treated" areas with "controled" ones or to include confounding variables that are measured at a prespecified geographic level. Alternatively, variables provided at certain geographic levels can be aggregated or averaged at higher ones—like calculating average ownership rates by municipality from ZIP

code data. The MAUP implies that one can get different estimates of treatment effects—or of the conditional covariances between variables—based on the geographic level of aggregation (Fotheringham and Wong, 1991).

In some cases, researchers do not know the exact extent and scope of treatment variation or the perceptions of locals about which areas constitute their neighborhoods. In these situations, it is sensible to ascertain the robustness of results to alternative geographic definitions. For instance, Moraga et al. (2019) create random rectangular neighborhoods to study native flight in Spain, finding their results to be robust to different square sizes.

In many cases, the researcher will be content with adhering to the one policy-relevant geographic definition. For instance, if a program is provided at the municipal level—or a variable changes discontinuously at municipal borders—then it is appropriate to use municipal outcomes as encountered in the real world.

16.2   Microdata Overkill

A related problem arises from the actual need for data disaggregation. At times, the promise of big data is overstated, as the relevant answer might be obtained simply by comparing conditional means. For instance, if one wishes to discern the propensity of a specific population group toward a particular behavior, a mere data tabulation may suffice. Acquiring and analyzing individual microdata or mapping the sub-population frequency vís-a-vís the behavior might not add significant value. However, microdata is critical if the covariances across variables are important and/or change with time and space. Microdata is also vital for model-driven applications, for understanding heterogeneity in behavior, or when there are nonlinearities or spillovers between variables and/or observational units (Blundell et al., 1993).

16.3   Time Series versus Cross-Section

A similar concern arises with aggregation across time periods. Real-time data might be superfluous when the treatment or the independent variables of interest do not change much, or high-frequency changes manifest with high signal-to-noise ratios. In the first scenario, retaining high-frequency data could be excessive and geographic cross-sections might be adequate. In the latter, models utilizing long differences in data might be more precise. For instance, Saiz (2007) examines the impact of immigration on metropolitan housing prices, contrasting the use of annual data on rents with "long" changes in the same variable, as measured by the decennial census.

16.4 Ecological "Fallacy"

The opposite problem arises when behavior measured at aggregated levels does not mirror individual behavior, especially when individual actions are the object of study. Suppose the average propensity of people in the "orange" group to vote for candidate A surpasses that of "green" individuals: $P(A/O)>P(A/G)$. Intuitively, a regression of the local share of orange individuals on the share of votes for A *by neighborhood* might yield an estimate of $P(A/O)$. However, orange individuals might lean toward voting for A if surrounded by other oranges, indicating a local agglomeration effect. A slightly more challenging problem arises when there are spillovers across groups. For instance, the greens might lean toward voting for A when the neighborhood share of oranges is very high, reflecting a social pressure or conformity effect. Both scenarios can be modeled in neighborhood-level regressions by using nonlinearities; distinguishing between the two, however, might require micro-individual or suitably tabulated data. Nevertheless, often the relevant policy parameter is genuinely at the aggregate level. For instance, the focus might be on predicting the increase in the share of votes for A in a neighborhood following a rise in the local share of orange individuals.

Self-selection into neighborhoods can also create a misalignment between aggregate group coefficients and individual ones. For example, the types of orange individuals settling in predominantly orange neighborhoods might be more radical than the average. Moreover, omitted variables might also lead to deviations: perhaps, orange individuals gravitate toward places where a third variable, like local TV coverage, increases the probability of voting for A. Note that these challenges are common in *any research design*. Specifically, they can also afflict models employing individual micro-measurements.

The term "ecological fallacy" encompasses traditional problems that, like any empirical work, might demand careful statistical modeling and interpretation. Unfortunately, this concept has acquired totemic significance for some and can be misunderstood as an impossibility theorem. Schwartz (1994) aptly termed this the "fallacy of ecological fallacy." Empirically, for many applications, spillovers might be negligible (Openshaw, 1984). This implies that, often, an increase in the preponderance of a behavior in aggregate spatial data as a group's share increases indicates that group members indeed exhibit such behavior disproportionately.

16.5 Interactions across Spatial Units

The often misinterpreted "ecological fallacy" highlights potential behavioral spillovers within units, but urban spatial data may also reflect spillovers or commonalities across neighboring units. The field of spatial econometrics deals with modeling these data generation processes

across spatial units (Anselin, 2010).

A direct form of spatial covariance arises from omitted variables exhibiting spatial correlation. For instance, neighboring observational units might be prone to presenting similar values of an omitted variable. Alternatively, the variance of the dependent variable or the behavioral parameters might be spatially correlated. Both scenarios emphasize the importance of appropriately clustering standard errors (Abadie et al., 2023).

Treatments can influence neighboring geographic areas, leading to spillover effects. Moreover, and more complexly, potential spillovers and feedback loops might impact the dependent variable. As an illustration, local governments may compete against proximate municipalities to offer better services, meaning that an influential determinant of a locality's expenditure could be the expenditure habits of its neighbors (Brueckner, 2003). Such contexts require careful modeling of the strategic interactions between units and the potential outcomes in general equilibrium.

16.6 Measurement Error

The common issue of measurement error might intensify in the context of spatial data. Beyond the challenge of inaccurately measuring variables, the geolocation of pertinent spatial units can also be misjudged. Inaccurate, biased, or incomplete data can lead to misguided conclusions (Ravi and Ravi, 2015).

The complications stemming from measurement errors become evident when considering potential pitfalls in satellite imagery. Factors such as weather conditions, time of day, and seasonality can adversely affect the image quality, leading to potential inaccuracies. Similarly, while nighttime imagery provides valuable insights, variations in brightness, due to different lighting technologies or light pollution levels, might lead to inconsistencies. Such inconsistencies can result in data interpretation errors and may even correlate spatially.

16.7 Sample Selection and Representativeness

While large datasets might seem comprehensive, they do not necessarily guarantee representativeness. Take, for instance, data from credit card usage. Its use may introduce sampling biases by capturing primarily transactions from particular demographic groups or income brackets—likely those of high socioeconomic status with good access to the financial system. Similarly, social media data might not always mirror the broader population since users predominantly self-select based on demographics. Furthermore, offline behavior may not always follow the same patterns as online interactions. This does not invalidate such datasets, but it does warrant caution in interpreting results and might need statistical modeling techniques to

correct for biases. One potential solution is reweighting data to better reflect the population of interest.

Complications worsen when data selection is self-based on unobservable factors. This problem may be more salient with crowdsourced data, where volunteer contributors may not behave the same way as the general population. In these cases, researchers should try to validate the representatives of their samples.

For instance, Saiz et al. (2018) express concerns about whether individuals uploading images to platforms like Google Maps or Flickr might have aesthetic preferences divergent from the broader population. To address this issue, they conduct a survey on building aesthetics, incorporating a question for online posting frequency. They find that the building ratings of frequent online posters are statistically identical to those of people who are not active online, thereby validating their crowdsourcing approach.

16.8 Perceptual and Cognitive Biases of Sources

Methods that rely on crowdsourced data or vast textual repositories may inadvertently inherit biases from those who contributed the data. This challenge is akin to issues in artificial intelligence, where models might unintentionally adopt the biases present in human-provided classifications (Ntoutsi et al., 2020).

16.9 Behavior-Altering Measurement

The act of observation can sometimes alter the behavior of those being observed, a phenomenon typically termed the "Hawthorne effect" (Adair, 1984; McCambridge et al., 2014). Such effects become more likely with sensor or CDR data, especially if individuals are aware they're being measured. While many phone applications notify users about data collection and research applications, the real awareness levels remain uncertain, given the long, often unread contracts.

16.10 Privacy

Privacy and data security concerns are important when working with microdata, such as individual credit card transactions. Rigorous anonymization, data safety, and disclosure protocols must be in place to protect individual privacy rights. Privacy issues encompass both technical (Samarati, 2001) and ethical (Richards and King, 2014) considerations. Confidentiality in the treatment of private information should be paramount for researchers. However, it is worth noting that the private sector is already using a trove of urban microdata. Lim-

iting data access only to those pursuing prosocial purposes, like research, may not be in society's best interest. Balancing privacy and broader societal benefits is critical (Tene and Polonetsky, 2011).

17 Conclusion

The introduction of large and diverse data sources has had a significant impact on urban science, geography, transportation, urban economics, and real estate research. These sources, ranging from geographic information maps to crowdsourced insights, have deepened our understanding of human behavior in cities. Their inclusion has transformed urban planning and business strategies, and has opened new possibilities for entrepreneurship and policy interventions. This change has also been crucial for the growth of smart city initiatives and highlights the rise of urban informatics as an interdisciplinary field combining traditional quantitative methods with new data-driven approaches.

Spatial data also open opportunities for research in applied economics. Our hope is that the sources and examples in this scoping paper can inspire further use and exploitation of these data typologies by labor economists. For instance, CDR data may help us better understand individual labor market outcomes and their relationships to family time use. Volumetric, building, and sensor data may help us explain the factors that drive labor productivity in the office.

While these data sources expand the scope of urban research, they also introduce challenges, ranging from issues in data collection to concerns about the replicability of findings. It is important for researchers to be aware of these challenges and maintain high standards in data analysis and interpretation. By doing this, we can make the most of the transformative power that these data sources offer and ensure they have tangible societal benefits.

References

Aaronson, Daniel, Daniel Hartley, and Bhashkar Mazumder, "The effects of the 1930s HOLC âredliningâ maps," *American Economic Journal: Economic Policy*, 2021, *13* (4), 355–392.

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge, "When should you adjust standard errors for clustering?," *The Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Abbiasov, Timur, Cate Heine, Edward L. Glaeser, Carlo Ratti, Sadegh Sabouri, Arianna Salazar-Miranda, and Paolo Santi, "The 15-Minute City Quantified Using Mobility Data," *SSRN Electronic Journal*, 2022.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez, "Automated linking of historical data," *Journal of Economic Literature*, 2021, *59* (3), 865–918.

Adair, John G, "The Hawthorne effect: a reconsideration of the methodological artifact.," *Journal of applied psychology*, 1984, *69* (2), 334.

Adams, Benjamin and Grant McKenzie, *Inferring Thematic Places from Spatially Referenced Natural Language Descriptions*

Agrawal, Ajay, Christian Catalini, and Avi Goldfarb, "Crowdfunding: Geography, social networks, and the timing of investment decisions," *Journal of Economics & Management Strategy*, 2015, *24* (2), 253–274.

Ahlfeldt, Gabriel and Alexandra Mastro, "Valuing iconic design: Frank lloyd wright architecture in Oak Park, Illinois," *Housing studies*, 2012, *27* (8), 1079–1099.

Akbar, Prottoy A, Victor Couture, Gilles Duranton, and Adam Storeygard, "The fast, the slow, and the congested: Urban transportation in rich and poor countries," *University of Pennsylvania*, 2022.

Akbar, Prottoy and Gilles Duranton, "Measuring the cost of congestion in highly congested city: Bogotá," 2017.

_ , Victor Couture, Gilles Duranton, and Adam Storeygard, "Mobility and congestion in urban India," *American Economic Review*, 2023, *113* (4), 1083–1111.

Alsina-Pagès, Rosa Ma, Pau Bergadà, and Carme Martínez-Suquía, "Changes in the soundscape of Girona during the COVID lockdown," *The Journal of the Acoustical Society of America*, 2021, *149* (5), 3416–3423.

Anagol, Santosh, Vimal Balasubramaniam, Tarun Ramadorai, and Antoine Uettwiller, "A Bad Bunch: Asset Value Under-Reporting in the Mumbai Real Estate Market," *Available at SSRN 4055401*, 2022.

Ananat, Elizabeth Oltmans, "The wrong side (s) of the tracks: The causal effects of racial segregation on urban poverty and inequality," *American Economic Journal: Applied Economics*, 2011, *3* (2), 34–66.

Anderson, Soren T and Sarah E West, "Open space, residential property values, and spatial context," *Regional Science and Urban Economics*, 2006, *36* (6), 773–789.

Ang, Li-Minn and Kah Phooi Seng, "Big sensor data applications in urban environments," *Big Data Research*, 2016, *4*, 1–12.

Anselin, Luc, "Thirty years of spatial econometrics," *Papers in regional science*, 2010, *89* (1), 3–25.

Arsanjani, Jamal Jokar, Christopher Barron, Mohammed Bakillah, and Marco Helbich, "Assessing the quality of OpenStreetMap contributors together with their contributions," in "Proceedings of the AGILE" 2013, pp. 14–17.

Asabere, Paul K, "The value of a neighborhood street with reference to the cul-de-sac," *The Journal of Real Estate Finance and Economics*, 1990, *3* (2), 185–193.

Athey, Susan, Billy Ferguson, Matthew Gentzkow, and Tobias Schmidt, "Estimating experienced racial segregation in US cities using large-scale GPS data," *Proceedings of the National Academy of Sciences of the United States of America*, 2021, *118* (46), e2026160118.

Balcan, Duygu, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases.," *Proceedings of the National Academy of Sciences of the United States of America*, dec 2009, *106* (51), 21484–21489.

Barajas, Jesus M and Anne Brown, "Not minding the gap: Does ride-hailing serve transit deserts?," *Journal of Transport Geography*, 2021, *90*, 102918.

Barkham, Richard, Sheharyar Bokhari, and Albert Saiz, "Urban big data: city management and real estate markets," *Artificial Intelligence, Machine Learning, and Optimization Tools for Smart Cities: Designing for Sustainability*, 2022, pp. 177–209.

Barr, Jason, Troy Tassier, and Rossen Trendafilov, "Depth to Bedrock and the Formation of the Manhattan Skyline, 1890–1915," *The Journal of economic history*, 2011, *71* (4), 1060–1077.

Barrington-Leigh, Chris and Adam Millard-Ball, "More connected urban roads reduce US GHG emissions," *Environmental Research Letters*, 2017, *12* (4).

Barrington-Leigh, Christopher and Adam Millard-Ball, "A century of sprawl in the United States," *Proceedings of the National Academy of Sciences of the United States of America*, 2015, *112* (27), 8244–8249.

Batty, Michael, "Big data, smart cities and city planning," *Dialogues in human geography*, 2013, *3* (3), 274–279.

_ , "Urban informatics and big data," *A Report to the ESRC Cities Expert Group*, 2013, pp. 1–36.

Baum-Snow, Nathaniel, "Did highways cause suburbanization?," *The quarterly journal of economics*, 2007, *122* (2), 775–805.

Ben-Joseph, Eran, "20 City Design in the Age of Digital Ubiquity," *Companion to Urban Design*, 2011.

Bendeck, Alexander and Clio Andris, *Text Mining and Spatial Analysis of Yelp Data to Support Socially Vibrant Cities*, Vol. 1, Association for Computing Machinery, 2022.

Berghauser Pont, M.Y.; Haupt, P.A., "Space Matrix: Space, density and urban form." PhD dissertation, Tu Delft 2009.

Berry, Robert, Caitlin Hafferty, Orford Scott, and Lucy Clarke, "Exploring the evolution of GIS research using bibliographic data," 2023.

Biljecki, Filip and Koichi Ito, "Street view imagery in urban analytics and GIS: A review," *Landscape and Urban Planning*, 2021, *215*, 104217.

Black, Sandra E, "Do better schools matter? Parental valuation of elementary education," *The quarterly journal of economics*, 1999, *114* (2), 577–599.

Blumenstock, Joshua E., "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda," *Information Technology for Development*, 2012, *18* (2), 107–125.

Blundell, Richard, Panos Pashardes, and Guglielmo Weber, "What do we learn about consumer demand patterns from micro data?," *The American Economic Review*, 1993, pp. 570–597.

Boarnet, Marlon G, Saksith Chalermpong, and Elizabeth Geho, "Specification issues in models of population and employment growth," *Papers in regional Science*, 2005, *84* (1), 21–46.

Boeing, Geoff, "Off the Grid. . . and Back Again?: The Recent Evolution of American Street Network Planning and Design," *Journal of the American Planning Association*, 2020, pp. 1–30.

_ , "Spatial information and the legibility of urban form: Big data in urban morphology," *International Journal of Information Management*, 2021, *56*.

Bolitzer, B and N R Netusil, "The impact of open spaces on property values in Portland, Oregon," *Journal of Environmental Management*, 2000, *59* (3), 185–193.

Bonczak, Bartosz and Constantine E. Kontokosta, "Large-scale parameterization of 3D building morphology in complex urban landscapes using aerial LiDAR and city administrative data," in "Computers, Environment and Urban Systems," Vol. 73 2019, pp. 126–142.

Bounie, David, Youssouf Camara, and John W. Galbraith, "Consumer mobility and expenditure during the COVID-19 containments: Evidence from French transaction data," *European Economic Review*, 2023, *151*.

Bourassa, Steven C, Martin Hoesli, and Jian Sun, "A simple alternative house price index method," *Journal of Housing Economics*, 2006, *15* (1), 80–97.

Boyarchenko, Nina, Andreas Fuster, and David O Lucca, "Understanding mortgage spreads," *The Review of Financial Studies*, 2019, *32* (10), 3799–3850.

Bracke, Philippe, "House prices and rents: Microevidence from a matched data set in Central London," *Real Estate Economics*, 2015, *43* (2), 403–431.

Bratu, Cristina, Oskari Harjunen, and Tuukka Saarimaa, "City-wide effects of new housing supply: Evidence from moving chains," *VATT Institute for Economic Research Working Papers*, 2021, *146*.

Brown, B B and V L Cropper, "New urban and standard suburban subdivisions: Evaluating psychological and social goals," *Journal of the American Planning Association*, 2001, *67* (4), 402–419.

Brueckner, Jan K, "Strategic interaction among governments: An overview of empirical studies," *International regional science review*, 2003, *26* (2), 175–188.

Burchfield, Marcy, Henry G Overman, Diego Puga, and Matthew A Turner, "Causes of sprawl: A portrait from space," *The Quarterly Journal of Economics*, 2006, *121* (2), 587–633.

Burke, Marshall, Anne Driscoll, David B Lobell, and Stefano Ermon, "Using satellite imagery to understand and promote sustainable development," *Science*, 2021, *371* (6535), eabe8628.

Cabrera, Joseph F and Jonathan C Najarian, "How the Built Environment Shapes Spatial Bridging Ties and Social Capital," *Environment and Behavior*, 2015, *47* (3), 239–267.

Calem, Paul, Jeanna Kenney, Lauren Lambie-Hanson, and Leonard Nakamura, "Appraising home purchase appraisals," *Real Estate Economics*, 2021, *49* (S1), 134–168.

Carlino, Gerald A and Albert Saiz, "Beautiful city: Leisure amenities and urban growth," *Journal of Regional Science*, 2019, *59* (3), 369–408.

Carvalho, Vasco M. and Hansen, Stephen, Álvaro Ortiz, Juan Ramon Garcia, Tomasa Rodrigo, Rodriguez Sevi Mora, and Pep Ruiz de Aguirre, "Tracking the COVID-19 Crisis

with High-Resolution Transaction Data (April 2020).," *CEPR Discussion Paper No. DP14642.*, 2020, (April).

Cervero, Robert and Kara Kockelman, "Travel demand and the 3Ds: Density, diversity, and design," *Transportation Research Part D: Transport and Environment*, sep 1997, *2* (3), 199–219.

Chambers, David, Christophe Spaenjers, and Eva Steiner, "The rate of return on real estate: Long-run micro-level evidence," *The Review of Financial Studies*, 2021, *34* (8), 3572–3607.

Charmes, Jacques, "Time use across the world: Findings of a world compilation of time use surveys," *Background Paper for Human Development Report*, 2015.

Charu, Vivek, Scott Zeger, Julia Gog, Ottar N Bjørnstad, Stephen Kissler, Lone Simonsen, Bryan T Grenfell, and Cécile Viboud, "Human mobility and the spatial transmission of influenza in the United States," *PLOS Computational Biology*, 2017, *13* (2), 1–23.

Chatterjee, Kiron, Samuel Chng, Ben Clark, Adrian Davis, Jonas De Vos, Dick Ettema, Susan Handy, Adam Martin, and Louise Reardon, "Commuting and wellbeing: a critical overview of the literature with implications for policy and future research," *Transport reviews*, 2020, *40* (1), 5–34.

Chen, Xi and William D Nordhaus, "Using luminosity data as a proxy for economic statistics," *Proceedings of the National Academy of Sciences*, 2011, *108* (21), 8589–8594.

Chetty, Raj, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter, "The opportunity atlas: Mapping the childhood roots of social mobility," Technical Report, National Bureau of Economic Research 2018.

_ , Nathaniel Hendren, Patrick Kline, and Emmanuel Saez, "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623.

Cho, Man and Isaac F Megbolugbe, "An empirical analysis of property appraisal and mortgage redlining," *The Journal of Real Estate Finance and Economics*, 1996, *13*, 45–55.

Choi, Hyunyoung and Hal Varian, "Predicting the present with Google Trends," *Economic record*, 2012, *88*, 2–9.

Ciccone, Antonio, "The Effect of WWI Military Casualties on the Population Distribution in Germany," 2021.

Ciepłuch, Błażej, Ricky Jacob, Peter Mooney, and Adam C Winstanley, "Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps," in "Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010" University of Leicester 2010, p. 337.

Clapp, John M and Carmelo Giaccotto, "Estimating price indices for residential property: a comparison of repeat sales and assessed value methods," *Journal of the American Statistical Association*, 1992, *87* (418), 300–306.

Connelly, Roxanne, Christopher J Playford, Vernon Gayle, and Chris Dibben, "The role of administrative data in the big data revolution in social science research," *Social science research*, 2016, *59*, 1–12.

Conzelmann, Claire; Salazar-Miranda, Arianna; Phan, Toan;Hoffman, Jeremy, "Long-term Effects of Redlining on Environmental Risk Exposure," 2023.

Conzen, M R G, *Alnwick, Northumberland: A Study in Town-Plan Analysis* number 27, Hassell Street Press, 1960.

Costa, Dora L, Matthew E Kahn, Christopher Roudiez, and Sven Wilson, "Persistent social networks: Civil war veterans who fought together co-locate in later life," *Regional science and urban economics*, 2018, *70*, 289–299.

Couture, Victor, Jonathan I. Dingel, Allison Green, Jessie Handbury, and Kevin R. Williams, "JUE Insight: Measuring movement and social contact with smartphone data: a real-time application to COVID-19," *Journal of Urban Economics*, 2022, *127* (202).

Damm, Anna Piil, "Ethnic enclaves and immigrant labor market outcomes: Quasi-experimental evidence," *Journal of Labor Economics*, 2009, *27* (2), 281–314.

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Sciences*, 2014, *111* (45), 15888–15893.

Di Clemente, Riccardo, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González, "Sequences of purchases in credit card data reveal lifestyles in urban populations.," *Nature communications*, aug 2018, *9* (1), 3330.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas, "More tweets, more votes: Social media as a quantitative indicator of political behavior," *PloS one*, 2013, *8* (11), e79449.

Doll, ChristopherN. H, Jan-Peter Muller, and Christopher D Elvidge, "Night-time Imagery as a Tool for Global Mapping of Socioeconomic Parameters and Greenhouse Gas Emissions," *AMBIO: A Journal of the Human Environment*, 2000, *29* (3), 157–162.

Donaldson, Dave and Adam Storeygard, "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, nov 2016, *30* (4), 171–198.

Dong, Xiaowen, Alfredo J. Morales, Eaman Jahani, Esteban Moro, Bruno Lepri, Burcin Bozkaya, Carlos Sarraute, Yaneer Bar-Yam, and Alex Pentland, "Segregated interactions in urban and online space," *EPJ Data Science*, 2020, *9* (1), 20.

Dougal, Casey, Christopher A Parsons, and Sheridan Titman, "Urban vibrancy and corporate growth," *The Journal of Finance*, 2015, *70* (1), 163–210.

Doyle, John, Peter Hung, Ronan Farrell, and Seán McLoone, "Population Mobility Dynamics Estimated from Mobile Telephony Data," *Journal of Urban Technology*, 2014, *21* (2), 109–132.

Drelichman, Mauricio, Jordi Vidal-Robert, and Hans-Joachim Voth, "The long-run effects of religious persecution: Evidence from the Spanish Inquisition," *Proceedings of the National Academy of Sciences*, 2021, *118* (33), e2022881118.

Ebener, Steeve, Christopher Murray, Ajay Tandon, and Christopher C. Elvidge, "From wealth to health: Modelling the distribution of income per capita at the sub-national level using night-time light imagery," *International Journal of Health Geographics*, 2005, *4*.

Edin, Per-Anders, Peter Fredriksson, and Olof Åslund, "Ethnic enclaves and the economic success of immigrantsâEvidence from a natural experiment," *The quarterly journal of economics*, 2003, *118* (1), 329–357.

Einav, Liran and Jonathan Levin, "Economics in the age of big data," *Science*, 2014, *346* (6210), 1243089.

Elvidge, Christopher D., Jeffrey Safran, Benjamin Tuttle, Paul Sutton, Pierantonio Cinzano, Donald Pettit, John Arvesen, and Christopher Small, "Potential for global mapping of development via a nightsat mission," *GeoJournal*, 2007, *69* (1-2), 45–53.

Ewing, Reid and Robert Cervero, "Travel and the Built Environment: A Meta-Analysis," *Journal of the American Planning Association*, jun 2010, *76* (3), 265–294.

Fairweather, Daryl, Matthew E Kahn, Robert D Metcalfe, and Sebastian Sandoval-Olascoaga, "The Impact of Climate Risk Disclosure on Housing Search and Buying Dynamics: Evidence from a Nationwide Field Experiment with Redfin," 2023.

Fan, Hongchao, Alexander Zipf, Qing Fu, and Pascal Neis, "Quality assessment for building footprints data on OpenStreetMap," *International Journal of Geographical Information Science*, 2014, *28* (4), 700–719.

Farrell, Diana, Lindsay Relihan, and Marvin Ward, "Going the distance: big data on resident access to everyday goods," *Available at SSRN 2966200*, 2017.

Farrell, Susan J, Tim Aubry, and Daniel Coulombe, "Neighborhoods and neighbors: Do they contribute to personal well-being?," *Journal of Community Psychology*, 2004, *32* (1), 9–25.

Feick, Rob and Colin Robertson, "A multi-scale approach to exploring urban places in geotagged photographs," *Computers, Environment and Urban Systems*, sep 2015, *53*, 96–109.

Fisher, Jeffrey D and David Geltner, "De-lagging the NCREIF index: transaction prices and reverse-engineering," *Real Estate Finance*, 2000, *17* (1), 7–22.

Foote, Christopher L and Paul S Willen, "Mortgage-default research and the recent foreclosure crisis," *Annual Review of Financial Economics*, 2018, *10*, 59–100.

Foth, Marcus, *Handbook of research on urban informatics: the practice and promise of the real-time city: the practice and promise of the real-time city*, IGI Global, 2008.

Fotheringham, A Stewart and David WS Wong, "The modifiable areal unit problem in multivariate statistical analysis," *Environment and planning A*, 1991, *23* (7), 1025–1044.

Fragkias, Michail and Karen C Seto, "Modeling Urban Growth in Data-Sparse Environments: A New Approach," *Environment and Planning B: Planning and Design*, oct 2007, *34* (5), 858–883.

Frank, L D and G Pivo, "Impacts of mixed use and density on utilization of three modes of travel: single-occupant vehicle, transit, and walking," *Transportation research record*, 1994, pp. 44–44.

Fuerst, Franz and Patrick McAllister, "Green noise or green value? Measuring the effects of environmental certification on office values," *Real estate economics*, 2011, *39* (1), 45–69.

Garcia-López, Miquel-Àngel, Adelheid Holl, and Elisabet Viladecans-Marsal, "Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities," *Journal of Urban Economics*, 2015, *85*, 52–67.

__ , **Jordi Jofre-Monseny, Rodrigo Martínez-Mazza, and Mariona Segú**, "Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona," *Journal of Urban Economics*, 2020, *119*, 103278.

Gauvin, Laetitia, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto, "Gender gaps in urban mobility," *Humanities and Social Sciences Communications*, 2020, *7* (1), 11.

Geertman, Stan, Joseph Ferreira, Robert Goodspeed, and John Stillwell, *Introduction to âplanning support systems and smart citiesâ*, Springer, 2015.

Geltner, David, "Bias in appraisal-based returns," *Real Estate Economics*, 1989, *17* (3), 338–352.

__ , "The Repeated Measures Regression-Based Index: A Better Way to Construct Appraisal-Based Indexes of Commercial Property Value," *Real Estate Finance*, 1996, *12*, 29–35.

__ , **Bryan D MacGregor, and Gregory M Schwann**, "Appraisal smoothing and price discovery in real estate markets," *Urban Studies*, 2003, *40* (5-6), 1047–1064.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Ghahramani, Mohammadhossein, Nadina J Galle, Fábio Duarte, Carlo Ratti, and Francesco Pilla, "Leveraging artificial intelligence to analyze citizens' opinions on urban green space," *City and Environment Interactions*, 2021, *10*, 100058.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, 2009, *457* (7232), 1012–1014.

Glaeser, Edward L, "Nowcasting the Local Economy : Using Yelp Data to Measure Economic Activity at Scale," 2018.

González, Marta C, César A Hidalgo, and Albert-László Barabási, "Understanding individual human mobility patterns," *Nature*, 2008, *453* (7196), 779–782.

Goodchild, Michael F, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, 2007, *69*, 211–221.

_ , "THE ROOTS OF GIS," *The Routledge Handbook of Geospatial Technologies and Society*, 2023.

Graells-Garrido, Eduardo, Oscar Peredo, and José García, "Sensing urban patterns with antenna mappings: The case of Santiago, Chile," *Sensors (Switzerland)*, 2016, *16* (7), 1–25.

Grannis, Rick, "The importance of trivial streets: Residential streets and residential segregation," *American Journal of Sociology*, 1998, *103* (6), 1530–1564.

Gu, Yizhen, Chang Jiang, Junfu Zhang, and Ben Zou, "Subways and road congestion," *American Economic Journal: Applied Economics*, 2021, *13* (2), 83–115.

Hahm, Yeankyoung, Heeyeun Yoon, and Yunwon Choi, "The effect of built environments on the walking and shopping behaviors of pedestrians; A study with GPS experiment in Sinchon retail district in Seoul, South Korea," *Cities*, 2019, *89*, 1–13.

Hajrasouliha, Amir and Li Yin, "The impact of street network connectivity on pedestrian volume," *Urban Studies*, 2015, *52* (13), 2483–2497.

Haklay, Mordechai, "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets," *Environment and planning B: Planning and design*, 2010, *37* (4), 682–703.

Hall, Peter Geoffrey, *Cities in civilization*, Vol. 21, Citeseer, 1998.

Haranko, Karri, Emilio Zagheni, Kiran Garimella, and Ingmar Weber, "Professional gender gaps across US cities," in "Proceedings of the International AAAI Conference on Web and Social Media," Vol. 12 2018.

Harari, Mariaflavia, "Cities in Bad Shape: Urban Geometry in India," *American Economic Review*, aug 2020, *110* (8), 2377–2421.

Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti, "Geo-located Twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science*, 2014, *41* (3), 260–271.

Heine, Cate, Cristina Marquez, Paolo Santi, Marcus Sundberg, Miriam Nordfors, and Carlo Ratti, "Analysis of mobility homophily in Stockholm based on social network data," *PLOS ONE*, 2021, *16* (3), 1–14.

_ , Kevin P O Keeffe, Paolo Santi, Li Yan, and Carlo Ratti, "Travel distance, frequency of return and the spread of disease," 2022, pp. 1–11.

Henao, Alejandro and Wesley E Marshall, "The impact of ride-hailing on vehicle miles traveled," *Transportation*, 2019, *46* (6), 2173–2194.

Henderson, J Vernon, Adam Storeygard, and David N Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, apr 2012, *102* (2), 994–1028.

Herath, Shanaka and Gunther Maier, "The hedonic price method in real estate and housing market research: a review of the literature," 2010.

Heroy, Samuel, Isabella Loaiza, Alex Pentland, and Neave OâClery, "Are neighbourhood amenities associated with more walking and less driving? Yes, but predominantly for the wealthy," *Environment and Planning B: Urban Analytics and City Science*, 2023, *50* (4), 958–982.

Hielkema, Hendrik and Patrizia Hongisto, "Developing the Helsinki smart city: The role of competitions for open data applications," *Journal of the Knowledge Economy*, 2013, *4*, 190–204.

Hill, Robert J, "Hedonic price indexes for residential housing: A survey, evaluation and taxonomy," *Journal of economic surveys*, 2013, *27* (5), 879–914.

Hillier, Bill and Julienne Hanson, *The Social Logic of Space*, Cambridge University Press, 1984.

_ and _ , *The social logic of space*, Cambridge: Cambridge University Press, 2005.

Hipp, John R and Andrew J Perrin, "The Simultaneous Effect of Social Distance and Physical Distance on the Formation of Neighborhood Ties," *City Community*, 2009, *8* (1), 5–25.

Hong, Jinhyun, David McArthur, and Varun Raturi, "Did safe cycling infrastructure still matter during a COVID-19 lockdown?," *Sustainability*, 2020, *12* (20), 8672.

Hortaçsu, Ali, F Asís Martínez-Jerez, and Jason Douglas, "The geography of trade in online transactions: Evidence from eBay and mercadolibre," *American Economic Journal: Microeconomics*, 2009, *1* (1), 53–74.

Hu, Yingjie, Chengbin Deng, and Zhou Zhou, "A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments," *Annals of the American Association of Geographers*, 2019, *109* (4), 1052–1073.

Imbens, Guido W and Donald B Rubin, "Rubin causal model," in "Microeconometrics," Springer, 2010, pp. 229–241.

Jardim, Ekaterina S, Mark C Long, Robert Plotnick, Emma van Inwegen, Jacob L Vigdor, and Hilary Wething, "Boundary Discontinuity Methods and Policy Spillovers," Technical Report, National Bureau of Economic Research 2022.

Jenkins, Andrew, Arie Croitoru, Andrew T Crooks, and Anthony Stefanidis, "Crowdsourcing a Collective Sense of Place," *PLOS ONE*, 2016, *11* (4), 1–20.

Jha, Akshaya and Andrea La Nauze, "US Embassy air-quality tweets led to global health benefits," *Proceedings of the National Academy of Sciences*, 2022, *119* (44), e2201092119.

Johnson, Erik B, Alan Tidwell, and Sriram V Villupuram, "Valuing curb appeal," *The Journal of Real Estate Finance and Economics*, 2020, *60* (1), 111–133.

Kang, Jian, Marco Körner, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu, "Building instance classification using street view images," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, *145*, 44–59.

Kang, Yuhao, Fan Zhang, Wenzhe Peng, Song Gao, Jinmeng Rao, Fabio Duarte, and Carlo Ratti, "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land Use Policy*, 2020, *111*, 104919.

Keele, Luke J and Rocio Titiunik, "Geographic boundaries as regression discontinuities," *Political Analysis*, 2015, *23* (1), 127–155.

Kholodilin, Konstantin A, Fernando A López, David Rey Blanco, and Pelayo Gonzalez Arbués, "Lessons from an aborted second-generation rent control in Catalonia," 2022.

Khosla, Aditya, Byoungkwon An An, Joseph J Lim, and Antonio Torralba, "Looking beyond the visible scene," in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition" 2014, pp. 3710–3717.

Kii, Masanobu, Tetsuya Tamaki, Tatsuya Suzuki, and Atsuko Nonomura, "Estimating urban spatial structure based on remote sensing data.," *Scientific reports*, may 2023, *13* (1), 8804.

Kishino, Yasue, Yoshinari Shirai, Koh Takeuchi, Takayuki Suyama, Futoshi Naya, and Naonori Ueda, "Regional garbage amount estimation and analysis using car-mounted motion sensors," in "Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers" 2018, pp. 110–113.

Kitchin, Rob, "The real-time city? Big data and smart urbanism," *GeoJournal*, 2014, *79*, 1–14.

Kok, Nils and Maarten Jennen, "The impact of energy labels and accessibility on office rents," *Energy policy*, 2012, *46*, 489–497.

Kolar, Vinay, Sayan Ranu, Anand Prabhu Subramainan, Yedendra Shrinivasan, Aditya Telang, Ravi Kokku, and Sriram Raghavan, "People in motion: Spatio-temporal analytics on Call Detail Records," *2014 6th International Conference on Communication Systems and Networks, COMSNETS 2014*, 2014, pp. 1–4.

Kreindler, Gabriel and Yuhei Miyauchi, "Measuring Commuting and Economic Activity Inside Cities with Cell Phone Records," *SSRN Electronic Journal*, 2021.

Krumme, Coco, Alejandro Llorente, Manuel Cebrian, Alex ("Sandy") Pentland, and Esteban Moro, "The predictability of consumer visitation patterns," *Scientific Reports*, 2013, *3* (1), 1645.

Kuchler, Theresa, Dominic Russel, and Johannes Stroebel, "JUE Insight: The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook," *Journal of Urban Economics*, 2022, *127*, 103314.

Kung, Kevin S., Kael Greco, Stanislav Sobolevsky, and Carlo Ratti, "Exploring universal patterns in human home-work commuting from mobile phone data," *PLoS ONE*, 2014, *9* (6).

Lai, Yuan and Constantine E Kontokosta, "Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities," *Computers, Environment and Urban Systems*, 2019, *78*, 101383.

Lang, William W and Leonard I Nakamura, "A model of redlining," *Journal of Urban Economics*, 1993, *33* (2), 223–234.

Laniado, David, Yana Volkovich, Salvatore Scellato, Cecilia Mascolo, and Andreas Kaltenbrunner, "The impact of geographic distance on online social interactions," *Information Systems Frontiers*, 2018, *20*, 1203–1218.

Law, Stephen, Brooks Paige, and Chris Russell, "Take a look around: Using street view and satellite images to estimate house prices," *ACM Transactions on Intelligent Systems and Technology*, 2019, *10* (5), 1–19.

Lazrak, Faroek, Peter Nijkamp, Piet Rietveld, and Jan Rouwendal, "The market value of cultural heritage in urban areas: an application of spatial hedonic pricing," *Journal of Geographical Systems*, 2014, *16*, 89–114.

Lenormand, Maxime, Thomas Louail, Oliva G Cantú-Ros, Miguel Picornell, Ricardo Herranz, Juan Murillo Arias, Marc Barthelemy, Maxi San Miguel, and José J Ramasco,

"Influence of sociodemographic characteristics on human mobility [corrected].," *Scientific reports*, may 2015, *5*, 10075.

Less, Elyse Levine, Patricia McKee, Traci Toomey, Toben Nelson, Darin Erickson, Serena Xiong, and Rhonda Jones-Webb, "Matching study areas using Google Street View: A new application for an emerging technology," *Evaluation and program planning*, 2015, *53*, 72–79.

Leyden, Kevin M., "Social Capital and the Built Environment: The Importance of Walkable Neighborhoods," *American Journal of Public Health*, 2003, *93* (9), 1546–1551.

Li, Han and Zhigang Li, "Road investments and inventory reduction: Firm level evidence from China," *Journal of Urban Economics*, 2013, *76*, 43–52.

Li, Linna, Michael F Goodchild, and Bo Xu, "Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr," *cartography and geographic information science*, 2013, *40* (2), 61–77.

Li, Xiaojiang and Carlo Ratti, "Mapping the spatial distribution of shade provision of street trees in Boston using Google Street View panoramas," *Urban Forestry & Urban Greening*, 2018, *31*, 109–119.

_ , **Paolo Santi, Theodore K. Courtney, Santosh K. Verma, and Carlo Ratti**, "Investigating the association between streetscapes and human walking activities using Google Street View and human trajectory data," *Transactions in GIS*, 2018, *22* (4), 1029–1044.

Li, Yan, Yiqun Chen, Abbas Rajabifard, Kourosh Khoshelham, and Mitko Aleksandrov, "Estimating building age from Google Street View images using deep learning," in "10th International Conference on Geographic Information Science" 2018.

Li, Yanchao, Ziyu Ran, Lily Tsai, and Sarah Williams, "Using call detail records to determine mobility patterns of different socio-demographic groups in the western area of Sierra Leone during early COVID-19 crisis," *Environment and Planning B: Urban Analytics and City Science*, 2023, *50* (5), 1298–1312.

Lin, Li, Liping Di, Chen Zhang, Liying Guo, and Yahui Di, "Remote sensing of urban poverty and gentrification," *Remote Sensing*, 2021, *13* (20), 1–20.

Linden, Leigh and Jonah E Rockoff, "Estimates of the impact of crime risk on property values from Megan's laws," *American Economic Review*, 2008, *98* (3), 1103–1127.

Lindenthal, Thies, "Beauty in the eye of the home-owner: Aesthetic zoning and residential property values," *Real Estate Economics*, 2020, *48* (2), 530–555.

_ **and Erik Barry Johnson**, "Machine Learning, Architectural Styles and Property Values," *SSRN Electronic Journal*, 2020, (April), 1–31.

Liu, Kang, Song Gao, Peiyuan Qiu, Xiliang Liu, Bo Yan, and Feng Lu, "Road2Vec: Measuring traffic interactions in urban road system from massive travel routes," *ISPRS International Journal of Geo-Information*, 2017, *6* (11), 7–13.

Liu, Ziyu, Anqi Yang, Mengyao Gao, Hong Jiang, Yuhao Kang, Fan Zhang, and Teng Fei, "Towards feasibility of photovoltaic road for urban traffic-solar energy estimation using street view image," *Journal of Cleaner Production*, 2019, *228*, 303–318.

Louail, Thomas, Maxime Lenormand, Oliva G. Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific Reports*, 2014, *4*, 1–12.

Louf, Rémi and Marc Barthelemy, "A typology of street patterns," *Journal of the Royal Society Interface*, 2014, *11* (101).

Lu, Tina and Ben Y Reis, "Internet search patterns reveal clinical course of COVID-19 disease progression and pandemic spread across 32 countries," *NPJ digital medicine*, 2021, *4* (1), 22.

Lu, Yi, Yiyang Yang, Guibo Sun, and Zhonghua Gou, "Associations between overhead-view and eye-level urban greenness and cycling behaviors," *Cities*, 2019, *88*, 10–18.

Lutzenhiser, Margot and Noelwah R. Netusil, "The effect of open spaces on a home's sale price," *Contemporary Economic Policy*, 2001, *19* (3), 291–298.

Maantay, Juliana, "Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research," *Health & place*, 2007, *13* (1), 32–56.

Mahtta, Richa, Anjali Mahendra, and Karen C. Seto, "Building up or spreading out? typologies of urban growth across 478 cities of 1 million+," *Environmental Research Letters*, 2019, *14* (12).

Malik, Praveen Kumar, Rajesh Singh, Anita Gehlot, Shaik Vaseem Akram, and Prabin Kumar Das, "Village 4.0: Digitalization of village with smart internet of things technologies," *Computers & Industrial Engineering*, 2022, *165*, 107938.

Malpezzi, Stephen et al., "Hedonic pricing models: a selective and applied review," *Housing economics and public policy*, 2003, *1*, 67–89.

Manaugh, Kevin, Luis F Miranda-Moreno, and Ahmed M El-Geneidy, "The effect of neighbourhood characteristics, accessibility, home–work location, and demographics on commuting distances," *Transportation*, 2010, *37*, 627–646.

Markou, Ioulia, Kevin Kaiser, and Francisco C Pereira, "Predicting taxi demand hotspots using automated Internet Search Queries," *Transportation Research Part C: Emerging Technologies*, 2019, *102*, 73–86.

Marshall, Stephen, *Streets patterns*, 1st ed ed., London ; New York: Spon, 2005.

Masdeval, Christian and Adriano Veloso, "Mining citizen emotions to estimate the urgency of urban issues," *Information systems*, 2015, *54*, 147–155.

Mast, Evan, "JUE Insight: The effect of new market-rate housing construction on the low-income housing market," *Journal of Urban Economics*, 2023, *133*, 103383.

McCambridge, Jim, John Witton, and Diana R Elbourne, "Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects," *Journal of clinical epidemiology*, 2014, *67* (3), 267–277.

McGreal, Stanley and Paloma Taltavull de La Paz, "An analysis of factors influencing accuracy in the valuation of residential properties in Spain," *Journal of Property Research*, 2012, *29* (1), 1–24.

Mitchell, Lewis, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth, "The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place," *PloS one*, 2013, *8* (5), e64417.

Mleczko, Matthew and Matthew Desmond, "Using natural language processing to construct a National Zoning and Land Use Database," *Urban Studies*, 2023.

Mooney, Stephen J, Katherine Wheeler-Martin, Laura M Fiedler, Celine M LaBelle, Taylor Lampe, Andrew Ratanatharathorn, Nimit N Shah, Andrew G Rundle, and Charles J DiMaggio, "Development and validation of a Google Street View pedestrian safety audit tool," *Epidemiology*, 2020, *31* (2), 301.

Moraga, Jesús Fernández-Huertas, Ada Ferrer i Carbonell, and Albert Saiz, "Immigrant locations and native residential preferences: Emerging ghettos or new communities?," *Journal of Urban Economics*, 2019, *112*, 133–151.

Moro, Esteban, Dan Calacci, Xiaowen Dong, and Alex Pentland, "Mobility patterns are associated with experienced income segregation in large US cities," *Nature communications*, 2021, *12* (1), 1–10.

Muller, Catherine L, Lee Chapman, CSB Grimmond, Duick T Young, and Xiaoming Cai, "Sensors and the city: a review of urban meteorological networks," *International Journal of Climatology*, 2013, *33* (7), 1585–1600.

Müller, Kalin, Charlotte Steinmeier, and Meinrad Küchler, "Urban growth along motorways in Switzerland," *Landscape and urban Planning*, 2010, *98* (1), 3–12.

Naik, Nikhil, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo, "Computer vision uncovers predictors of physical urban change," *Proceedings of the National Academy of Sciences*, 2017, *114* (29), 7571–7576.

Nasar, Jack L. and David A. Julian, "The psychological sense of community in the neighborhood," *Journal of the American Planning Association*, 1995, *61* (2), 178–184.

Ning, Huan, Zhenlong Li, Xinyue Ye, Shaohua Wang, Wenbo Wang, and Xiao Huang, "Exploring the vertical dimension of street view image based on deep learning: a case study on lowest floor elevation estimation," *International Journal of Geographical Information Science*, 2021, *35* (12), 1–26.

Niu, Haifeng and Elisabete A Silva, "Crowdsourced data mining for urban activity: Review of data sources, applications, and methods," *Journal of Urban Planning and Development*, 2020, *146* (2), 04020007.

Noonan, Douglas S, "Neighbours, barriers and urban environments: Are things' different on the other side of the tracks'?," *Urban Studies*, 2005, *42* (10), 1817–1835.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis et al., "Bias in data-driven artificial intelligence systemsâAn introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020, *10* (3), e1356.

Nyerges, Timothy, Robert B McMaster, and Helen Couclelis, "The SAGE handbook of GIS and society," *The SAGE Handbook of GIS and Society*, 2011, pp. 1–576.

Obradovich, Nick, Ömer Özak, Ignacio Martín, Ignacio Ortuño-Ortín, Edmond Awad, Manuel Cebrián, Rubén Cuevas, Klaus Desmet, Iyad Rahwan, and Ángel Cuevas, "Expanding the measurement of culture with a sample of two billion humans," *Journal of the Royal Society Interface*, 2022, *19* (190), 20220085.

O'grady, Trevor, "Spatial Institutions in Urban Economies: How City Grids Affect Density and Development," 2014.

Openshaw, Stan, "Ecological fallacies and the analysis of areal census data," *Environment and planning A*, 1984, *16* (1), 17–31.

Palchykov, Vasyl, Marija Mitrović, Hang-Hyun Jo, Jari Saramäki, and Raj Kumar Pan, "Inferring human mobility using communication patterns.," *Scientific reports*, aug 2014, *4*, 6174.

Paldino, Silvia, Dániel Kondor, Iva Bojic, Stanislav Sobolevsky, Marta C. González, and Carlo Ratti, "Uncovering urban temporal patterns from geo-tagged photography," *PLoS ONE*, 2016, *11* (12), 1–14.

Papinski, Dominik, Darren M Scott, and Sean T Doherty, "Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS," *Transportation research part F: traffic psychology and behaviour*, 2009, *12* (4), 347–358.

Perry, Joe N and Philip M Dixon, "A new method to measure spatial association for ecological count data," *Ecoscience*, 2002, *9* (2), 133–141.

Polgreen, Philip M, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein, "Using internet searches for influenza surveillance," *Clinical infectious diseases*, 2008, *47* (11), 1443–1448.

Ponce-Lopez, Roberto and Joseph Ferreira, "Identifying and characterizing popular non-work destinations by clustering cellphone and point-of-interest data," *Cities*, 2021, *113*, 103158.

Prabhu, R., B. Parvathavarthini, and R. A. Alagu Raja, "Slum Extraction from High Resolution Satellite Data using Mathematical Morphology based approach," *International Journal of Remote Sensing*, 2021, *42* (1), 172–190.

Preis, Tobias, Federico Botta, and Helen Susannah Moat, "Sensing global tourism numbers with millions of publicly shared online photographs," *Environment and Planning A: Economy and Space*, 2020, *52* (3), 471–477.

Qiu, Waishan, Ziye Zhang, Xun Liu, Wenjing Li, Xiaojiang Li, Xiang Xu, and Xiaokai Huang, "Subjective or objective measures of street environment, which are more effective in explaining housing prices?," *Landscape and Urban Planning*, 2022, *221*, 104358.

Quercia, Daniele, Neil O'Hare, and Henriette Cramer, "Aesthetic Capital: What Makes London Look Beautiful, Quiet, and Happy ?," *Proceedings of the 17th ACM conference on Computer supported cooperative work social computing*, 2014, pp. 945–955.

Ramani, Arjun and Nicholas Bloom, "The Donut effect of COVID-19 on cities," Technical Report, National Bureau of Economic Research 2021.

Ratti, Carlo, *Smart city, smart citizen*, EGEA spa, 2013.

_ , **Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams**, "Mobile landscapes: Using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, 2006, *33* (5), 727–748.

Ravi, Kumar and Vadlamani Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, 2015, *89*, 14–46.

Reades, Jonathan, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti, "Cellular Census: Explorations in Urban Data Collection," *IEEE Pervasive Computing*, 2007, *6* (3), 30–38.

Reddy, Alla, Alex Lu, Santosh Kumar, Victor Bashmakov, and Svetlana Rudenko, "Entry-only automated fare-collection system data used to infer ridership, rider destinations, unlinked trips, and passenger miles," *Transportation research record*, 2009, *2110* (1), 128–136.

Relihan, Lindsay, "Is online retail killing coffee shops? estimating the winners and losers of online retail using customer transaction microdata," *Estimating the Winners and Losers of Online Retail Using Customer Transaction Microdata (March 1, 2022). London School of Economics CEP Discussion Paper*, 2022, (1836).

Richards, Neil M and Jonathan H King, "Big data ethics," *Wake Forest L. Rev.*, 2014, *49*, 393.

Rico-Juan, Juan Ramón and Paloma Taltavull de La Paz, "Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain," *Expert Systems with Applications*, 2021, *171*, 114590.

Rogers, George and Sineenart Sukolratanametee, "Neighborhood design and sense of community: Comparing suburban neighborhoods in Houston Texas," *Landscape and Urban Planning*, 2009, *92*, 325–334.

Rolheiser, Lyndsey A and Chengzhen Dai, "Beyond density: municipal expenditures and the shape and location of development," *Urban Geography*, 2019, *40* (8), 1097–1123.

Rosen, Sherwin, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of political economy*, 1974, *82* (1), 34–55.

Rosenthal, Stuart S and William C Strange, "The attenuation of human capital spillovers," *Journal of Urban Economics*, 2008, *64* (2), 373–389.

Saiz, Albert, "Immigration and housing rents in American cities," *Journal of urban Economics*, 2007, *61* (2), 345–371.

— , "The Geographic Determinants of Housing Supply," *Quarterly Journal of Economics*, 2010, *125* (3), 1253–1296.

— , "The Global Housing Affordability Crisis: Policy Options and Strategies," *MIT Center for Real Estate Research Paper*, 2023, (23/01).

— **and Arianna Salazar-Miranda**, "Real trends: The future of real estate in the united states," *MIT Center for Real Estate Research Paper*, 2017, (5).

— **and Luyao Wang**, "Physical geography and traffic delays: Evidence from a major coastal city," *Environment and Planning B: Urban Analytics and City Science*, 2023, *50* (1), 218–243.

— **and Susan Wachter**, "Immigration and the neighborhood," *American Economic Journal: Economic Policy*, 2011, *3* (2), 169–188.

— **and Uri Simonsohn**, "Proxying for unobservable variables with internet document-frequency," *Journal of the European Economic Association*, 2013, *11* (1), 137–165.

_ , **Arianna Salazar-Miranda, and James Bernard**, "Crowdsourcing architectural beauty: Online photo frequency predicts building aesthetic ratings," *PLOS ONE*, 2018, *13* (7), 1–15.

Salazar-Miranda, Arianna, "The shape of segregation: The role of urban form in immigrant assimilation," *Cities*, 2020, *106*, 102852.

Salazar-Miranda, Arianna, "The micro persistence of layouts and design: quasi-experimental evidence from the United States housing corporation," *Regional Science and Urban Economics*, 2022, *95*, 103755.

_ , "The Role of Neighborhood Design for Social and Environmental Sustainability: Evidence From US Cities," 2022, (September).

_ , **Cate Heine, Fabio Duarte, Katja Schechtner, and Carlo Ratti**, "Measuring the impact of slow zones on street life using social media," *Cities*, 2022, *131*, 104010.

_ , **Guangyu Du, Claire Gorman, Fabio Duarte, Washington Fajardo, and Carlo Ratti**, "Favelas 4D: Scalable methods for morphology analysis of informal settlements using terrestrial laser scanning data," *Environment and Planning B: Urban Analytics and City Science*, 2022, *49* (9), 2345–2362.

Salazar-Miranda, Arianna, Zhuangyuan Fan, Fabio Duarte, and Carlo Ratti, "Desirable streets: Using deviations in pedestrian trajectories to measure the value of the built environment," *Computers, Environment and Urban Systems*, 2021, p. 101563.

Samarati, Pierangela, "Protecting respondents identities in microdata release," *IEEE transactions on Knowledge and Data Engineering*, 2001, *13* (6), 1010–1027.

Sánchez-Aparicio, María, Susana Del Pozo, Jose Antonio Martín-Jiménez, Enrique González-González, Paula Andrés-Anaya, and Susana Lagüela, "Influence of lidar point cloud density in the geometric characterization of rooftops for solar photovoltaic studies in cities," *Remote Sensing*, 2020, *12* (22), 1–21.

Sant'Anna, Vinicios and Gustavo Cortes, "Send Them Back? The Real Estate Consequences of Repatriations," *The Center for Growth and Opportunity*, 2023.

Saraiva, Miguel, Irina MatijošaitienÄ, Saloni Mishra, and Ana Amante, "Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics," *ISPRS International Journal of Geo-Information*, 2022, *11* (7).

Schaefer, Karl A, "The effect of floodplain designation/regulations on residential property values: a case study in North York, Ontario," *Canadian Water Resources Journal*, 1990, *15* (4), 319–332.

Schauer, Lorenz, Martin Werner, and Philipp Marcus, "Estimating crowd densities and pedestrian flows using wi-fi and bluetooth," in "Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services" 2014, pp. 171–177.

Schuetz, Jenny, Vicki Been, and Ingrid Gould Ellen, "Neighborhood effects of concentrated mortgage foreclosures," *Journal of Housing Economics*, 2008, *17* (4), 306–319.

Schwartz, Sharon, "The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences.," *American journal of public health*, 1994, *84* (5), 819–824.

Seiferling, Ian, Nikhil Naik, Carlo Ratti, and Raphäel Proulx, "Green streets- Quantifying and mapping urban trees with street-level imagery and computer vision," *Landscape and Urban Planning*, 2017, *165*, 93–101.

Seresinhe, Chanuki Illushka, Helen Susannah Moat, and Tobias Preis, "Quantifying scenic areas using crowdsourced data," *Environment and Planning B: Urban Analytics and City Science*, 2017, p. 026581351668730.

_ , **Tobias Preis, and Helen Susannah Moat**, "Using deep learning to quantify the beauty of outdoor places," *Royal Society Open Science*, 2017, *4* (7), 170170.

Seto, Karen C. and Michail Fragkias, "Quantifying spatiotemporal patterns of urban land-use change in four cities of China with time series landscape metrics," *Landscape Ecology*, 2005, *20* (7), 871–888.

Seto, Karen C, Burak Güneralp, and Lucy R Hutyra, "Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools," *Proceedings of the National Academy of Sciences*, 2012, *109* (40), 16083–16088.

_ , **Michail Fragkias, Burak Güneralp, and Michael K Reilly**, "A Meta-Analysis of Global Urban Land Expansion," *PLOS ONE*, 2011, *6* (8), 1–9.

Shen, Chen and Ping Li, "Green housing on social media in China: A text mining analysis," *Building and Environment*, 2023, *237*, 110338.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Lee, Jacob Carlson, and Weining Li, "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis," *International Conference on Document Analysis and Recognition*, 2021, pp. 131–146.

Shertzer, Allison, Tate Twinam, and Randall P Walsh, "Race, ethnicity, and discriminatory zoning," *American Economic Journal: Applied Economics*, 2016, *8* (3), 217–246.

Shimizu, Chihiro, Kazuo G Nishimura, and Tsutomu Watanabe, "House prices from magazines, realtors, and the Land Registry," *BIS Paper*, 2012, (64f).

Sikder, Amit Kumar, Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, Kemal Akkaya, and Mauro Conti, "IoT-enabled smart lighting systems for smart cities," in "2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)" IEEE 2018, pp. 639–645.

Siksna, Arnis, "City centre blocks and their evolution: A comparative study of eight American and Australian CBDs," *Journal of Urban Design*, oct 1998, *3* (3), 253–283.

Singh, Vivek Kumar, Burcin Bozkaya, and Alex Pentland, "Money Walks: Implicit Mobility Behavior and Financial Well-Being," *PLOS ONE*, 2015, *10* (8), 1–17.

Sirmans, Stacy, David Macpherson, and Emily Zietz, "The composition of hedonic pricing models," *Journal of real estate literature*, 2005, *13* (1), 1–44.

Smith-Clarke, Christopher, Afra Mashhadi, and Licia Capra, "Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks," *Conference on Human Factors in Computing Systems - Proceedings*, 2014, pp. 511–520.

Sobolevsky, Stanislav, Izabela Sitko, Remi des Combes, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti, "Money on the Move: Big Data of Bank Card Transactions as the New Proxy for Human Mobility Patterns and Regional Delineation. The Case of Residents and Foreign Visitors in Spain," in "2014 IEEE International Congress on Big Data" 2014, pp. 136–143.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási, "Limits of Predictability in Human Mobility," *Science*, 2010, *327* (5968), 1018–1021.

Song, Yan and Gerrit-Jan Knaap, "New urbanism and housing values: a disaggregate assessment," *Journal of Urban Economics*, 2003, *54* (2), 218–238.

Soto, Víctor and Enrique Frías-Martínez, "Automated land use identification using cell-phone records," *MobiSys'11 - Compilation Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services and Co-located Workshops - HotPlanet'11*, 2011, pp. 17–22.

Stacy, Christina, Chris Davis, Yonah Slifkin Freemark, Lydia Lo, Graham MacDonald, Vivian Zheng, and Rolf Pendall, "Land-use reforms and housing costs: Does allowing for increased density lead to greater affordability?," *Urban Studies*, 2023.

Sun, Maoran, Changyu Han, Quan Nie, Jingying Xu, Fan Zhang, and Qunshan Zhao, "Understanding building energy efficiency with administrative and emerging urban big data by deep learning in Glasgow," *Energy and Buildings*, 2022, *273*, 112331.

Sutton, P, D Roberts, C Elvidge, and K Baugh, "Census from Heaven: An estimate of the global human population using night-time satellite imagery," *International Journal of Remote Sensing*, 2001, *22* (16), 3061–3076.

Sutton, Paul C., "A scale-adjusted measure of "Urban Sprawl" using nighttime satellite imagery," *Remote Sensing of Environment*, 2003, *86* (3), 353–369.

Tajima, Kayo, "New estimates of the demand for urban green space: Implications for valuing the environmental benefits of Bostonâs big dig project," *Journal of Urban affairs*, 2003, *25* (5), 641–655.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman, "Geography of Twitter networks," *Social networks*, 2012, *34* (1), 73–81.

Tene, Omer and Jules Polonetsky, "Privacy in the age of big data: a time for big decisions," *Stan. L. Rev. Online*, 2011, *64*, 63.

Teo, Hoong Chen, Tze Kwan Fung, Xiao Ping Song, Richard N. Belcher, Kelly Siman, Ian Z.W. Chan, and Lian Pin Koh, "Increasing contribution of urban greenery to residential real estate valuation over time," *Sustainable Cities and Society*, 2023, *96* (May), 104689.

Thakuriah, Piyushimita, Nebiyou Y Tilahun, and Moira Zellner, *Introduction to seeing cities through big data: Research, methods and applications in urban informatics*, Springer, 2017.

Tizzoni, Michele, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M. Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C. González, and Vittoria Colizza, "On the Use of Human Mobility Proxies for Modeling Epidemics," *PLoS Computational Biology*, 2014, *10* (7).

Traud, Amanda L, Peter J Mucha, and Mason A Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics and its Applications*, 2012, *391* (16), 4165–4180.

Twinam, Tate, "The long-run impact of zoning: Institutional hysteresis and durable capital in Seattle, 1920–2015," *Regional Science and Urban Economics*, 2018, *73*, 155–169.

Vanky, Anthony P., Santosh K. Verma, Theodore K. Courtney, Paolo Santi, and Carlo Ratti, "Effect of weather on pedestrian trip count and duration: City-scale evaluations using mobile phone application data," *Preventive Medicine Reports*, 2017, *8*, 30–37.

Vetrò, Antonio, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, and Federico Morando, "Open data quality measurement framework: Definition and application to Open Government Data," *Government Information Quarterly*, 2016, *33* (2), 325–337.

Villeneuve, Hannah and William O'Brien, "Listen to the guests: Text-mining Airbnb reviews to explore indoor environmental quality," *Building and Environment*, 2020, *169*, 106555.

Wang, Qi, Nolan Edward Phillips, Mario L Small, and Robert J Sampson, "Urban mobility and neighborhood isolation in Americaâs 50 largest cities," *Proceedings of the National Academy of Sciences*, 2018, *115* (30), 7735–7740.

Williams, Sarah, Adam White, Peter Waiganjo, Daniel Orwa, and Jacqueline Klopp, "The digital matatu project: Using cell phones to create an open source data for Nairobi's semi-formal bus system," *Journal of Transport Geography*, 2015, *49*, 39–51.

Wilson, Bev and Cong Cong, "Beyond the supply side: Use and impact of municipal open data in the US," *Telematics and Informatics*, 2021, *58*, 101526.

Witte, Astrid De, Joachim Hollevoet, Frédéric Dobruszkes, Michel Hubert, and Cathy Macharis, "Linking modal choice to motility: A comprehensive review," *Transportation Research Part A: Policy and Practice*, 2013, *49*, 329–341.

Wong, Grace, "Has SARS infected the property market? Evidence from Hong Kong," *Journal of Urban Economics*, 2008, *63* (1), 74–95.

Wong, Maisy, "Estimating the distortionary effects of ethnic quotas in Singapore using housing transactions," *Journal of Public Economics*, 2014, *115*, 131–145.

Wood, Lisa, Lawrence D Frank, and Billie Giles-Corti, "Sense of community and its relationship with walking and neighborhood design," *Social Science Medicine*, 2010, *70* (9), 1381–1390.

Wurm, Michael, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, and Hannes Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2019, *150*, 59–69.

Xu, Yang, Alexander Belyi, Paolo Santi, and Carlo Ratti, "Quantifying segregation in an integrated urban physical-social space," *Journal of the Royal Society Interface*, 11 2019, *16* (160), 20190536.

Yabe, Takahiro, Bernardo García Bulle Bueno, Xiaowen Dong, Alex Pentland, and Esteban Moro, "Behavioral changes during the COVID-19 pandemic decreased income diversity of urban encounters," *Nature Communications*, 2023, *14* (1).

Yang, Juncheng, Helena Rong, Yuhao Kang, Fan Zhang, and Andrea Chegut, "The Financial Impact of Street-Level Greenery on New York Commercial Buildings," *Available at SSRN 3714858*, 2020.

Yang, Liya, Carlos Marmolejo Duarte, and Pablo Martí Ciriquián, "Quantifying the relationship between public sentiment and urban environment in Barcelona," *Cities*, 2022, *130*, 103977.

Yin, Hang and Christian Berger, "When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets," in "2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)" IEEE 2017, pp. 1–8.

Yoshimura, Yuji, Stanislav Sobolevsky, Juan N Bautista Hobin, Carlo Ratti, and Josep Blat, "Urban association rules: Uncovering linked trips for shopping behavior," *Environment and Planning B: Urban Analytics and City Science*, 2018, *45* (2), 367–385.

Yu, Shi-Ming, Sun-Sheng Han, and Chee-Hian Chai, "Modeling the value of view in high-rise apartments: A 3D GIS approach," *Environment and Planning B: Planning and Design*, 2007, *34* (1), 139–153.

Yun, Hae Young, Christopher Zegras, and Daniel Heriberto Palencia Arreola, "âDigitalizing Walkabilityâ: Comparing Smartphone-Based and Web-Based Approaches to Measuring Neighborhood Walkability in Singapore," *Journal of Urban Technology*, 2019, *26* (3), 3–43.

Zandbergen, Paul A, "A comparison of address point, parcel and street geocoding techniques," *Computers, Environment and Urban Systems*, 2008, *32* (3), 214–232.

Zhang, Fan, Arianna Salazar-Miranda, Fábio Duarte, Lawrence Vale, Gary Hack, Yu Liu, Michael Batty, and Carlo Ratti, "Urban Visual Intelligence: Studying Cities with AI and Street-level Imagery," 2023, pp. 1–41.

_ , **Ding Zhang, Yu Liu, and Hui Lin**, "Representing place locales using scene elements," *Computers, Environment and Urban Systems*, 2018, *71*, 153–164.

_ , **Lun Wu, Di Zhu, and Yu Liu**, "Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2019, *153*, 48–58.

Zhang, Kai, Min Chen, Yue Yang, Teng Zhong, Rui Zhu, Fan Zhang, Zhen Qian, Guonian Lü, and Jinyue Yan, "Quantifying the photovoltaic potential of highways in China," *Applied Energy*, 2022, *324*, 119600.

Zhang, Liying, Tao Pei, Xi Wang, Mingbo Wu, Ci Song, Sihui Guo, and Yijin Chen, "Quantifying the urban visual perception of Chinese traditional-style building with street view images," *Applied Sciences*, 2020, *10* (17), 5963.

Zhao, Zhan, Haris N Koutsopoulos, and Jinhua Zhao, "Identifying hidden visits from sparse call detail record data," *Transactions in Urban Data, Science, and Technology*, 2022, *1* (3-4), 121–141.

Zhou, Yuyu, Xuecao Li, Wei Chen, Lin Meng, Qiusheng Wu, Peng Gong, and Karen C Seto, "Satellite mapping of urban built-up heights reveals extreme infrastructure gaps and inequalities in the Global South," *Proceedings of the National Academy of Sciences*, 2022, *119* (46), e2214813119.

Zhu, Shanjiang and David Levinson, "Do people use the shortest path? An empirical test of wardrop's first principle," *PLoS ONE*, 2015, *10* (8), 1–18.

Ziersch, Anna M, Fran E Baum, Colin MacDougall, and Christine Putland, "Neighbourhood life and social capital: the implications for health," *Social Science Medicine*, 2005, *60* (1), 71–86.

Zou, Shengyuan and Le Wang, "Detecting individual abandoned houses from Google Street View: A hierarchical deep learning approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, *175*, 298–310.