7

What Types of Physical and Built Environment Can We Find in Digital Data?

Lindsey Smith

Abstract

Processes such as urbanization demonstrate how human activity influences the physical environment and the subsequent implications for Earth's natural systems. Correspondingly, changes to different environments, and in particular built environments, are linked with human behavior and health. Understanding these relationships requires the definition and measurement of environments. Considering advancements in the collection and processing of high-volume and high-velocity geospatial data, this chapter seeks to outline features of physical and built environments that may be identified from digital data. Attention is given to open data with varying spatial and temporal resolutions. Administrative data, remote sensing imagery, and data from stationary sensors provide contextual information such as the rate of urban expansion and changes in air quality. Mobile and social sensing enable the collection of high-resolution data that contribute to the identification of smaller-scale features. Developments in classification techniques, such as deep learning, provide the opportunity to explore human–environment interactions in real time. Although challenges exist related to data integration and categorization and must be resolved by future research, the combination of data from multiple sources adds value and holds promise for improving our understanding of the patterns that rapidly change landscapes, and the role of environments in shaping human behavior.

Introduction

Urbanization has modified Earth's surface and contributed to considerable changes in land cover and land use. Understanding how humans use and alter the landscape has long been important for disaster risk planning and sustainable resource management purposes (Meyer and Turner 1992; Turner et al. 2007). While human activity has influenced the form and quality of both the

physical and built environments through processes such as urbanization, deforestation, and agriculture, a bidirectional association exists whereby the environment simultaneously shapes human behaviors and health.

Socioecological models recognize multiple drivers of behavior (Sallis et al. 2006). Alongside social and interpersonal factors, spatial and temporal factors related to the environments with which people interact are now embraced as determinants of behavior and health (Rainham et al. 2010). Conditions and opportunities vary by city and neighborhoods, producing social and health inequalities (Marmot 2005; Santana 2017). As a potentially modifiable target for intervention, the built environment has been increasingly addressed by research and policy. For example, a wealth of urban health literature has explored air pollution, water contamination, food environments, green spaces, and active travel infrastructure as well as their relationships with related behavioral and health outcomes (Brunekreef and Holgate 2002; Houlden et al. 2018; Lytle and Sokol 2017; Van Holle et al. 2012). Health, well-being, responsible production, as well as sustainable and resilient cities feature in the United Nations Sustainable Development Goals as strategic approaches for tackling inequality.

Understanding different environments and their relationships with human activity is therefore important. These relationships are, however, complex in nature and operate through multiple mediators and operators (Dahlgren and Whitehead 1991; Rutter et al. 2017). To date, much research has been limited by a focus on single attributes of the environment, narrow or simplistic conceptualizations of space, and either measures of risk exposure or behavioral change without consideration for how these interact (Frank et al. 2019). Advances in geospatial and computational technologies have contributed to the emergence of big spatial data. For example, remote sensing, geographic information systems (GIS), and global positioning systems (GPS) enable the collection of data with high spatial and temporal resolution. This, in turn, creates new opportunities to characterize environments and enhance understanding of human–environment relationships.

This chapter outlines features of the physical and built environments that have important implications for social science research. Digital data sources that enable the identification of such features are subsequently discussed. Focus is given to big, open datasets that are accessible to researchers and may be used to create comparable and scalable measures. Approaches developed for processing data, however, may also hold relevance for individual-level information, such as the collection of imagery from wearable cameras in cohort studies. For each data type, an overview and examples will be provided, along with a discussion of strengths and limitations. In closing, opportunities and key challenges that pertain to all data sources are highlighted to guide discussions on how measures and frameworks may be developed to advance future research.

Physical and Built Environments

The physical environment refers to physical surroundings such as air, geological and climate conditions, water, vegetation, and the built environment. The built environment, more specifically, encompasses spaces and places that have been created or modified by humans to support human needs and activities. This ranges in scale from cities to neighborhoods to infrastructure and features of urban form such as buildings and urban parks.

Table 7.1 provides a sampling of physical and built environments. The concepts discussed in this chapter are contingent on the author's research in urban spaces in high-income countries such as the United Kingdom and Canada (see also Appendix 7.1). A different research scenario may require additional or alternative concepts not listed.

Physical environments that occur on Earth's surface may be described by land cover type such as vegetation (e.g., forest, grassland, cropland), water types (e.g., wetland, open water), urban area, ice, bare soil, and rock. These environments may be further categorized by land use (i.e., the purpose for which land is utilized by people). Areas with the same land cover type can have different land uses, which may be influenced by geographical factors including the availability of resources, existing infrastructure, and proximity to urban populations. A range of land use categories have been identified and studied across a number of disciplines. Common land use types studied within an urban context include residential, commercial, transportation, recreational, and institutional. These uses, and how they change over time, provide information for planning and may influence the types and spatial configuration of built environment features that are developed. The built environment may be subject to administrative boundaries and notions of access or ownership. Features of the built environment include transport infrastructure and services such as roads, footpaths, and health-care facilities. Measures of features (e.g., the density of intersections, retail outlets, and residential units) may also be used to derive information about the value of spaces in relation to human–environment interactions, such as a walkability score.

Environments may be considered at a range of scales that correspondingly affect the type and frequency of human behavior associated with them. Macroscale environments, such as heat and rainfall, may contribute to hazardous events and affect displacement, migration, and food production. Microscale environments, such as food retailers and facilities designed for physical activity in the built environment, may encourage specific and more regular behaviors, such as types of food purchased and modes of travel used. Environments, particularly at the microscale, are moderated by quality. For instance, a park close to a busy road may experience higher rates of air and noise pollution, affecting pathways to use and associated health and well-being outcomes. While quality of environments may have an objective measure in the data, factors

Table 7.1 Examples of physical and built environments and corresponding open digital data sources. Both spatial scale and population impact decrease from top to bottom of the table.

Environments	Digital Data Type and Example Source
Climate and weather: season, temperature, precipitation, natural hazard event	*Remote sensing*: global coverage of surface temperature *Stationary sensory*: *in situ* weather recordings *Social media*: citizen response to flood event
Land cover: vegetation, water, soil, urban	*Remote sensing*: Classification of land cover from multispectral satellite sensors
Land use: agricultural, conservation, residential, commercial, industrial, institutional	*Administrative data*: food and agriculture, business registry, census population data *Social media*: spatiotemporal clustering of users
Boundaries: protected areas, municipalities, plots, buildings	*Administrative data*: census boundary files, land information
Features: *Greenspace*: parks, gardens, trees *Bluespace*: harbors, lakes, rivers, water features *Transport infrastructure*: roads, footpaths, cycle lanes, bus stops *Services*: health care, education, leisure, housing, retail *Utilities*: wastewater system, power station and lines	*Administrative data*: digitized land survey data *Participatory sensing*: volunteered points of interest plotted through open-source platform *Mobile sensing*: street view imagery of store fronts
Quality: traffic, air quality, noisescape, lighting, litter, human perceptions	*Stationary sensory*: estimated surface models of air pollution from sensor network *Mobile sensing*: street view imagery of building damage *Participatory sensing*: mobile crowdsensing of noise *Social media:* semantic analysis of geotagged tweets

such as safety, cleanliness, and noisescape can be perceived or experienced uniquely by different groups and individuals.

Digital Data Sources

In social science research, environmental data may be quantified by a geographical unit (e.g., point location, administrative boundary, address buffer, activity space) to measure exposure, access, change, or use of space. Resultant

metrics may be subsequently linked to social data (e.g., based on home or work address) for analysis.

Administrative Data

Spatial data representing the physical and built environments may be obtained from existing datasets. Administrative data, such as land survey data or census data, are typically derived from organizations and institutions. Although administrative data has largely been neglected from the discussions associated with big data, such data can provide reliable information at national scales (Connelly et al. 2016). In contrast to raw sensing data, administrative data are usually cleaned (e.g., inaccuracies and inconsistencies in the data are rectified) and organized by data specialists into a format available for download and use within a GIS.

The ESRI open data hub provides access to over 210,000 open GIS datasets that have been collected from organizations around the world. Searches of key features and areas return related web maps, live dashboards, and datasets. Data can be downloaded in a variety of formats, including vector formats such as Shapefile and GeoJSON, which are commonly used for representing geographic features as points, lines, and polygons, and raster formats such as GeoTIFF that stores geospatial information as grids of pixels (see also Brinkhoff, this volume). Each dataset includes metadata describing the data source and the date when data were most recently updated. Often, local governments provide access to regional vector files representing infrastructure, land use, and municipal facilities through an open data portal. The features available as well as the temporal and spatial range of these data may, however, be limited. Separate files for individual characteristics, such as parks, schools, and transport infrastructure, can make it difficult to map cities and spaces fully. More systematic examples of administrative data collected at the national level include DMTI Spatial (Digital Mapping Technologies Inc.) data in Canada and Ordnance Survey data in the United Kingdom; both are available to education institutions in their respective countries. These repositories enable consistent matching of environmental characteristics to national cohorts with geographical heterogeneity, such as the U.K. Biobank dataset (Sarkar et al. 2015), and novel analytical approaches that incorporate a combination of environmental characteristics at scale (Smith et al. 2019b).

In addition to spatial data files of vector features representing specific environmental characteristics, aggregated information can also describe environments by geographic units. Statistics Canada, for instance, provides annual and biannual information on greenness, parks, and trees on properties at national, provincial, and metropolitan levels. As recorded in the Canadian census, which

is updated every five years, population density also provides a proxy for residential density by dissemination area, census tract,[1] or larger regions.

Ultimately, administrative data can provide independent measures of the environment that range in type and scale, as well as information on groups or areas that may not be represented in social sensing data (see below). Further, administrative data are less likely to contain processing and measurement errors compared with those collected from human input or sensing (Groen 2012). While time and cost associated with data production may be reduced, the spatial coverage and availability of data may be uneven, both within and between countries, and the temporal frequency of data updates is often slow and may be inconsistent across datasets of the same area.

Sensor Data

In addition to administrative data, data collected from sensors provide a valuable source for deriving measures of the physical and built environment. Below, four primary types of sensing data are outlined: remote sensing, stationary sensing, mobile sensing, and social sensing.

Remote Sensing

Remote sensing provides information about the Earth's surface based on reflected or emitted radiation. Information is recorded by instruments at a distance, typically aboard a satellite or aircraft, and can be processed subsequently to identify features in the physical and built environment (Read and Torrado 2009). Key advantages of satellite data include its global and relatively long temporal coverage, which allows patterns and impacts of the global landscape to be systematically captured (Wulder et al. 2019). Correspondingly, satellite data have long been used to monitor land surface temperature, meteorological and climate conditions, and greenness, as well as to map and detect change in large-scale land cover such as water bodies, vegetation, bare soil, and urban infrastructure.

Reflective of the range of applications, a multitude of satellite-based remote sensing instruments exist (Horning 2019). Depending on age and purpose, sensors vary in spectral, spatial, and temporal resolution, which can affect image quality and accuracy of object detection. A characteristic example of satellites used to monitor land surface is the Landsat program, which was first launched in 1972. Since 2008, Landsat data have been available for download at no cost due to a change in data policy; this has contributed to an expansion of scientific studies (Hemati et al. 2021; Zhu et al. 2019). The most recent Landsat satellite,

[1] In the Canadian census, Statistics Canada uses geographic units such as dissemination areas (i.e., the smallest standard area available with a population of 400 to 700 persons) and census tracts (i.e., areas in large urban centers with a population of 2,500 to 8,000).

Landsat 9, carries the Operational Land Instrument (OLI) and the Thermal Infrared Sensor (TIRS), which permits multispectral images to be produced at a spatial resolution of 30 m. Compared with the coarse resolution of early land cover maps (300 m to 1 km), resultant data have improved land cover classification accuracy (Chen et al. 2015; Gong et al. 2013) and have been utilized at local, national, and global scales.

Studies utilizing Landsat data have explored common drivers of land cover change, including deforestation, urbanization, human activities, and abrupt events such as wildfires (Hemati et al. 2021). For example, a loss of 1.5 million km^2 in global forest cover was recorded between 2000 and 2012 (Hansen et al. 2013), and urban land cover in China was reported to have doubled between 1990 and 2010, replacing existing cropland (Wang et al. 2012). Such findings are substantiated by alternative remote sensing sources such as MODIS data (collected from NASA's Moderate Resolution Imaging Spectroradiometer sensors) (Schneider et al. 2010) and nighttime light observations used to identify urban clusters (collected from The Defense Meteorological Program Operational Line-Scan System) (Liu et al. 2012; Zhou et al. 2018).

Processing raw remote sensing data and, particularly, identifying land cover types requires the application of machine-learning algorithms and specialist knowledge of spectral classification. Techniques including support vector machine, random forest, decision tree and artificial neural networks have been increasingly used to classify land cover (Talukdar et al. 2020). The development and availability of global land cover products, such as FROM-GLC10 (Gong et al. 2019), allow users to access classified data without having to perform any data processing (Li et al. 2020). These products are, however, often limited in terms of their temporal coverage. As an intermediary approach, advances in cloud computing platforms such as Google Earth Engine have reduced the need for storage requirements and provide access to myriad large-scale datasets and algorithms for image processing (Gorelick et al. 2017).

While high-resolution remote sending data can aid the process of monitoring climate conditions and mapping land cover and changes, inherent constraints of the data limit classification accuracy (Talukdar et al. 2020). For example, difficulties arise in distinguishing subtle variations in vegetation types with similar spectral reflectance and small-scale features such as parking lots or small residential structures can be challenging to classify due to limited spatial resolution of Landsat imagery. Furthermore, information about land use (e.g., use of forestry for conservation or use of urban spaces for residential purposes) cannot be inferred, particularly in urban environments where single spaces may be used in multiple ways. Integrating remote sensing data with complementary sources such as administrative or social sensing data may therefore be important for improving accuracy, understanding how people interact with environments, and identifying finer-resolution built environments relevant to social science research (Yin et al. 2021).

Stationary Sensors

In contrast to the large-scale coverage of data collected by remote sensing, stationary sensors (e.g., environmental sensors and cameras) collect high-frequency information from single locations. They are suited for detecting daily changes in features and quality of smaller-scale environments.

Monitoring sites in cities are commonplace for observing traffic flows, travel modes, monitoring weather conditions, and detecting urban air quality and noise. Data from dynamic sensor streams may be broadcast for the purposes of real-time visualization (e.g., the Toronto ESRI live stream dashboard, which reports on transit, traffic, weather, and air quality for the metropolitan area). Alternatively, historic data collected at each sensor may be downloaded for analysis. Example applications include studies in the United Kingdom that utilize data collected hourly from fixed weather stations (Meteorological Office Integrated Data Archive System) to explore urban heat effects for climate change resilience (Emmanuel and Krüger 2012; Heaviside et al. 2015) and weather effects on mobility (Brum-Bastos et al. 2018). Researchers acknowledge, however, the limitations of sparse spatial coverage of stationary sensors.

To estimate exposure between networks of monitoring sites, surfaces such as Weather Research and Forecasting, dispersion, or land use regression models may be developed within a GIS. As part of the European Study of Cohorts for Air Pollution Effects project (ESCAPE), a land use regression model was generated using data from up to 80 passive samplers at 36 sites across Europe (Beelen et al. 2013). Additional predictors of land use, traffic, and geographic characteristics from administrative datasets were input into the model to derive average annual concentrations of particulate matter with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$), nitrogen dioxide (NO_2), and nitrogen oxides (NO_X) as continuous variables, enabling the attribution of value to the home addresses of participants in multiple cohort studies across Europe.

Despite the ability to estimate exposure from surface models, low density networks of monitoring points limit the capacity of models to capture accurately the spatial variability in concentrations being measured (Marshall et al. 2008). To address this and improve spatial and temporal variability, studies increasingly incorporate mobile (Deville Cavellin et al. 2016) and crowdsourced social sensor data, such as citizen weather station networks (Brousse et al. 2022) and microphone-enabled smartphone apps, to measure ambient noise levels (Marjanovic et al. 2017).

Mobile Sensors

Mobile and portable sensors have contributed to increased spatial coverage of sensor networks. Street view datasets, such as Google Street View (GSV), Bing Streetside, and Tencent (specific to China where these is no official

coverage of GSV) are examples of mobile sensor data that capture small-scale features in the built environment. GSV, the largest of these datasets, has full or partial coverage in 102 countries. Data are collected as 360° panoramic images from vehicles and updated (at most) annually, depending on location and urbanicity. Images are available for download via the GSV application programming interface (API) which provides access to the most recent imagery. The GSV "Time Machine" function and open-source packages (e.g., the module for downloading photos from GSV) further enable users to view and access historic data to assess retrospectively environmental change (Cândido et al. 2018; Cohen et al. 2020).

The emergence of street view imagery has proven useful for identifying features such as retail outlets and validating existing GIS datasets. In addition, fine image resolution has facilitated the identification of visual factors that affect the quality of the built environment, such as greenness, broken windows, potholes, property damage, litter, and the estimation of urban canyons based on sky openness and building height. Street imagery also makes it possible to identify "nudge factors," such as the presence of billboards advertising junk food (Egli et al. 2019; Huang et al. 2020), and relative perceptions of environmental quality across space. For example, comparison of GSV imagery with hand-drawn maps revealed Latin American schoolchildren were more aware of litter in natural compared with urban environments (De Veer et al. 2022). Lastly, street view imagery makes it possible to explore human interactions with the environment, such as the number of street users and their modes of travel (Goel et al. 2018; Ibrahim et al. 2021).

Identifying features at scale requires the application of deep learning techniques whereby models are first trained on a large sample of images. Applied examples include the use of semantic segmentation and convolutional neural networks to predict human perceptions of images (Zhang et al. 2018) as well as to identify streetscape green and blue spaces and to examine relationships with behavior and health outcomes, such as depression in the elderly in Beijing, China (Helbich et al. 2019), or walking in Hong Kong (Lu 2018). These complex approaches rely on pixel-level classification to recognize and understand the subtle differences of features within an urban scene.

Mobile sensor data are more cost- and time-effective than field audits for identifying visible environmental features. Previous studies report accurate and consistent agreements between field audits and the use of street view imagery, highlighting its potential for filling in missing information from stationary sensor and administrative datasets. Critiques include irregular spatial coverage and variable collection frequency. In the case of street view data, only a snapshot of locations is provided, and this does not capture dynamics and flows of urban spaces or account for differences by time of day, day of the week, or season. In addition, data coverage may be biased toward more commercial streets, given the focus on businesses within Google Maps. As with remote

sensing and administrative datasets, value is added to mobile sensor data when integrated with complementary information such as social sensing data.

Social Sensing

Social sensing involves the collection, processing, and analysis of crowdsourced data from humans using devices (Pandharipande 2021). Given the proliferation of smartphone usage with GPS and camera capabilities as well as the unprecedented use of social media platforms for broadcasting information, social sensing has received attention as a means to acquire data about cities and human–environment interactions at scale (Aggarwal and Abdelzaher 2013; Wang et al. 2015). Consequently, social sensing offers potential for applications in urban planning, transport, health, and crime prevention.

Social sensing can be categorized into (a) participatory social sensing, where participants are recruited or voluntarily contribute data about a geographical area, (b) the use of social media data, where the user is not purposively generating data to map environment features. Although not discussed here, population-level GPS data from smartphones also provides useful information for traffic flows and crowding.

Participatory Social Data. In general, the process of participatory sensing involves citizens voluntarily and intentionally uploading local information through a platform or application. This instance of user-generated content, coined volunteered geographic information (VGI) by Goodchild (2007), has enabled collaborative mapping of environmental features based on local knowledge of participants worldwide. Citizens have become empowered to collect and map features that may not traditionally be included in administrative datasets, such as cycling facilities and wheelchair routes.

One prominent example of VGI is OpenStreetMap (OSM), which has around 37,000 active contributors per month. OSM is an openly accessible and editable map of the world; contributors typically input points of interest (e.g., retail outlets, education and health facilities, transit stops) and linear features (e.g., rivers, roads, bus routes) that can be downloaded via various repositories. Key strengths of OSM include its community input and global coverage, yet concerns have been raised over biases in mapped features and the validity of data. As a result, researchers have sought to demonstrate reasonable comparisons with administrative datasets in specific regions of the world (Dorn et al. 2015; Haklay 2010), and tools such as TagInfo have been developed to guide users and encourage consistency when tagging features. Attempts to standardize OSM tags also enables them to be mapped to standard codes, such as the North American Industry Classification System, allowing for linkage and comparison with administrative data.

In addition to existing platforms for logging volunteered information, geotagging campaigns may utilize mobile crowdsensing to build a denser network

and more reliable measures of environmental conditions. For example, in Brazil, *Guiaderodas* (a technology company that promotes accessibility in built environments) relied on crowdsourcing to evaluate the accessibility of over 250,000 establishment locations for wheelchair users in over 115 countries. Data are subsequently used to provide information for app users to plan accessible routes. Measures of noise have also been recorded for studies using microphones on personal smartphones. Capturing more complex measures of weather and air pollution, for example, may require the use of specialist devices, which can limit the number of contributors (Brousse et al. 2022; Marjanovic et al. 2017).

Social Media Data. Social media platforms such as Twitter/X, Facebook, Instagram, Flickr, YouTube, online blogs, and review ratings sites allow billions of users to generate and share data in the form of text, image, or video. The use of social media on smartphones can also provide detailed contextual information, such as location and time, based on GPS.

Geotagged social media data may be downloaded through an API. Although not initially intended to provide environmental information, spatial and temporal clustering of check-in activities and geotagged tweets have been used to infer land use (Soliman et al. 2017; Zhan et al. 2014) and quality of parks by incorporating semantic content analysis (Kovacs-Györi et al. 2018). The potential of using a framework to integrate image, text, and maps has also been demonstrated in the context of a localized event: the release of water from flood control reservoirs in Houston during Hurricane Harvey in 2017 (Fan et al. 2020). A graph-based approach was first used to detect critical tweets, then an image-ranking algorithm for selecting relevant images, and lastly a kernel density estimation of geotagged locations was used to map the geographic coverage of disruptions. The combination of social media data types and approaches may therefore contribute to enhanced real-time situational awareness of rapid environmental changes, such as wildfires and flooding, as well as slower changes, such as evolving perceptions and definitions of land use (e.g., use of residential spaces for employment which accelerated during the COVID pandemic).

While social media data provide new opportunities for understanding human–environment interactions at scale, key limitations lie in its reliability and representativeness. It is difficult to infer the validity of information in text, and data remain biased toward social media users, specifically those who enable geo-location services. Only 1–2% of tweets are geotagged, calling for geocoding and geoparsing methods to extract additional locations (Middleton et al. 2018). In addition, concerns around geoprivacy due to the disclosure of individuals' sensitive locations have been raised. As a result, spatial data may need to be masked or aggregated to protect individuals from being identified through their location records.

Key Considerations

Below, areas that require further discussion are highlighted to guide the application of big data in environment–behavior research.

Data Integration

Each data source is associated with unique strengths and capabilities for identifying environmental features. For example, remote sensing imagery provides global coverage but cannot capture small-scale features or land use. Street view imagery provides greater resolution but often at lower temporal frequency, whereas social sensing can capture high-frequency information but data quality is limited. Selecting a single data source may be appropriate for identifying a single environmental feature; there will likely be trade-offs, however, in spatial and temporal coverage, and data quality. Furthermore, human behavior is embedded in a complex system of places, times, and environments. Much of the literature exploring relationships between the environment, behavior, and health has focused on single features. While useful for identifying associations with specific outcomes (e.g., walkability and walking), environmental characteristics coexist and interrelate. Reflecting on the growing recognition of the broader determinants of behavior and health, a more holistic and integrative approach to measuring environments is required if we are to gain a better understanding of these complex interactions.

Combining digital data sources enables multiple environments and outcomes of varying scales to be explored: from broad city-level influences on population health to feature-level influences on more personal behaviors. Curated data libraries provide the first step in bringing spatial data about the physical and built environment from diverse sources into a single location. Subsequent consideration must therefore be given to how data are integrated, particularly given different formats, time frames for collecting data, and disparate scales and coverage. Here, deep learning may provide an opportunity to bridge gaps between different data types (Zhang et al. 2019).

Data Categorization

Linking environmental data with information related to social, behavioral, and health outcomes creates possibilities for analyzing associations and exploring inequalities by place. The unit at which data are aggregated and categorized, however, may have implications for causality.

Sociodemographic, social, or health data, such as that acquired from the census or social media, may be aggregated to an administrative boundary such as a census tract. Matching data with environmental features quantified within the same unit enables broad population-level patterns to be observed. Such

analyses are limited by the modifiable areal unit problem, whereby the chosen spatial unit differentially impacts results. For example, the same data aggregated by census tract, postal code area, or an individual's neighborhood may yield different effects. Ecological fallacy (i.e., inferences made about individuals from group data) may also rise when investigating features of the built environment (Houston 2014). Exposure is considered to be the same for all who live in the same administrative unit, irrespective of mobility patterns and transport opportunities. Linking individual-level data from cohort studies helps to overcome this. Still, much work in this area has focused solely on the home address by quantifying features within a residential buffer. In doing so, researchers are at risk of the "uncertain geographic context problem" as relevant environments beyond the home neighborhood where behaviors occur are missed (Kwan 2012). Increasingly, studies use GPS data to capture more relevant spaces. Features are quantified within individuals' activity spaces, based on locations they have visited over time. Delineation of activity spaces has been inconsistent and studies have conflated measures of access with use of space (Smith et al. 2019a).

Consideration needs to be given, therefore, to the quantification and categorization of data to ensure its conceptual relevance for meeting study aims and enabling comparisons. Here, metadata can help ensure that data are not only findable but described transparently in terms of collection processes and applicable for previous use cases. As high-volume location data become increasingly available, consideration must be given to how spatial and temporal sequence patterning may be incorporated into measures, beyond simple delineations of activity spaces (Fuller and Stanley 2019). As measures begin to reflect environments of importance better, researchers must not lose sight of causal thinking and strive to develop stronger evidence on the pathways that act to influence use of spaces and changes in behavior.

Reproducibility

Given the volume of digital data and application of machine-learning methods to process data, it is important that methods are reported transparently. Code sharing sites such as GitHub allow researchers to test, collaborate, and build upon existing approaches. This has implications for replication, scalability, and comparison in future studies.

Conclusions

Taken together, the wealth and quality of openly available digital data enables the identification of a range of physical and built environments at different spatial and temporal scales. Remote sensing imagery, administrative data, and stationary sensors provide contextual information such as the rate of urban

expansion, changes in air pollution, and coverage of green spaces. Meanwhile, methodological advances in mobile and social sensing enable the collection and analysis of highly granular longitudinal data on small-scale features through VGI. Developments in classification techniques, such as deep learning algorithms, also permit real-time behaviors to be explored in place through social media and mobile devices. Compared with traditional field audits or the collection of information from study participants, acquiring and processing digital data can be much less resource intensive. This holds promise for improving our understanding of the patterns that are rapidly changing global and local landscapes, and the role of environments in shaping human behavior and health over time.

While the volume and velocity of openly available data may not yet match that of commercial or privatized data, the availability and variety of open data for characterizing physical and built environments is continually increasing. As computational capacity and data expand, users must provide key considerations as to (a) the representativeness and relevancy of data and (b) to integration, categorization, and reproducibility. Value is added when variable data from multiple sources are combined to explore spatial patterns or develop immediate strategies, such as the direction of humanitarian aid following a natural hazard event. Approaches to data preparation and analysis also have implications for causality, findings, and potential inferences. Transparency in communicating methods and findings, with efforts toward reproducibility, is therefore key to ensuring integrity and reliability in research.

Appendix 7.1: Explanation of Useful Terms

Big data: High-volume and velocity data which may be too large to be processed with traditional software applications. May be analyzed to reveal patterns, trends, and associations, especially relating to human behavior.

Open data: Freely available data that may be downloaded and modified.

Crowdsourced data: Contribution of information from a large number of people.

Geographic information system (GIS): Computer system for creating, storing, and analyzing spatial data.

Application programming interface (API): Intermediary software that enables the transmission of data between two applications.

Machine learning: A type of artificial intelligence (AI) that finds and learns from patterns in big data.

Deep learning: A type of machine learning that uses multiple layers of processing to find smaller patterns in big data.

Semantic image segmentation: Computer vision task in which each pixel of an image is labeled with a corresponding class of what is being represented.

Convolutional neural networks: A form of deep learning which uses multiple layers to process arrays of data such as those in images, and extract features.

Semantic analysis: Process of finding meanings in text.

This is a section of doi:10.7551/mitpress/15532.001.0001

Digital Ethology

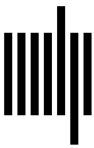
Human Behavior in Geospatial Context

Edited by: Tomáš Paus, Hye-Chung Kum

Citation:

Digital Ethology: Human Behavior in Geospatial Context
Edited by: Tomáš Paus, Hye-Chung Kum
DOI: 10.7551/mitpress/15532.001.0001
ISBN (electronic): 9780262378840
Publisher: The MIT Press
Published: 2024

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

Series Editor: J. R. Lupp
Editorial Assistance: A. Gessner, C. Stephen
Lektorat: BerlinScienceWorks

The book was set in TimesNewRoman and Arial.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-262-54813-7

10 9 8 7 6 5 4 3 2 1