Data Practices for Studying the Impacts of Environmental Amenities and Hazards with Nationwide Property Data

Christoph Nolte *Assistant Professor, Department of Earth and Environment, Boston University, Massachusetts; Affiliate Assistant Professor, Faculty of Computing and Data Sciences, Boston University, Massachusetts; chrnolte@bu.edu*

Kevin J. Boyle *Professor, Department of Agricultural and Applied Economics, Virginia Polytechnic Institute and State University, Blacksburg; Willis Blackwood Director, Blackwood Program in Real Estate, Virginia Polytechnic Institute and State University, Blacksburg; kjboyle@vt.edu*

Anita M. Chaudhry *Professor, Department of Economics, California State University, Chico; achaudhry@csuchico.edu*

Christopher Clapp *Assistant Instructional Professor, Harris School of Public Policy, University of Chicago, Illinois; cclapp@uchicago.edu*

Dennis Guignet *Assistant Professor, Department of Economics, Appalachian State University, Boone, North Carolina; guignetdb@appstate.edu*

Hannah Hennighausen *Assistant Professor, Department of Economics, University of Alaska, Anchorage; hbhennighausen@alaska.edu*

Ido Kushner *Research Assistant, Department of Earth and Environment, Boston University, Massachusetts; idok@bu.edu*

Yanjun Liao *Fellow, Resources for the Future, Washington, D.C.; yliao@rff.org*

Saleh Mamun *Postdoctoral Associate, Department of Applied Economics, University of Minnesota, St. Paul; Postdoctoral Associate, Natural Resources Research Institute, University of Minnesota, Duluth; salmamun@d.umn.edu*

Adam Pollack *Graduate Researcher, Department of Earth and Environment, Boston University, Massachusetts; Postdoctoral Research Associate, Dartmouth College, Hanover, New Hampshire; Adam.B.Pollack@dartmouth.edu*

Jesse Richardson *Professor, College of Law, West Virginia University, Morgantown; jesse.richardson@mail.wvu.edu*

Shelby Sundquist *Research Assistant, Department of Earth and Environment, Boston University, Massachusetts; Graduate Researcher, School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff; ss3988@nau.edu*

Kristen Swedberg *Graduate Student, Department of Agricultural and Applied Economics, Virginia Polytechnic Institute and State University, Blacksburg; ORISE Fellow, Office of Water, U.S. Environmental Protection Agency, Washington, D.C.; swedkm@vt.edu*

Johannes H. Uhl *Postdoctoral Researcher, Institute of Behavioral Science, University of Colorado, Boulder; Postdoctoral Researcher, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder; johannes.uhl@colorado.edu*

ABSTRACT *We discuss data quality and modeling issues inherent in the use of nationwide property data to value environmental amenities. By example of Zillow's Transaction and Assessment Database, a real estate database covering the United States, we identify challenges and propose guidance for (1) identifying arm's-length sales; (2) geolocating parcels and buildings; (3) identifying temporal links between transaction, assessor, and parcel data; (4) identifying property types, such as single-family homes and vacant lands; and (5) dealing with missing or mismeasured data for standard housing attributes. We review*

Land Economics • February 2024 • 100 (1): 200–221
DOI:10.3368/le.100.1.102122-0090R
ISSN 0023-7639; E-ISSN 1543-8325

current practice and show that how researchers address these issues can meaningfully influence research findings. (JEL Q51)

1. Introduction

Recent years have seen a rapid growth in empirical studies that use large-scale real estate data to value environmental amenities and hazards in the United States. This growth in empirical work is not new. The output of research papers using hedonic property value methods has been trending upward for three decades (Hanley and Czajkowski 2019), reflecting enduring interest in estimating peoples' environmental preferences, federal mandates to consider such values in regulatory cost-benefit analyses, and recent advances in computational capacity, econometric methods, and best practices (Bishop et al. 2020). A noteworthy recent trend is the publication of many hedonic analyses that make inferences across large and diverse geographic areas (see examples in Appendix Table A1). Although there are many sources of real estate microdata (e.g., state-level databases, private data aggregators, multiple listing services), an important contributor to this recent growth has been the decision of Zillow, a U.S. online real estate marketplace company, to share its Transaction and Assessment Database (ZTRAX) for free with U.S. academic, nonprofit, and government researchers between 2016 and 2023 (Zillow Group 2021b).

Access to large-scale real estate data has many potential benefits for economic research. It can help researchers improve and expand the set of available estimates of people's preferences for various amenities associated with property locations and characteristics (Bernstein, Gustafson, and Lewis 2019; Clarke and Freedman 2019; Albouy, Christensen, and Sarmiento-Barbieri 2020; Baldauf, Garlappi, and Yannelis 2020; Murfin and Spiegel 2020). It can narrow gaps in the geographic coverage of evidence derived from small-scale studies (Guignet et al. 2022). And it can mitigate some risks identified as contributing to a "credibility crisis" in environmental and resource economics (Ferraro and Shukla 2020, 2022). For instance, broader access to nationwide data puts more researchers in a position to reproduce and replicate findings from prior studies and test their generalizability across different contexts. This greatly increases the credibility of existing results and the intellectual merit of such findings (Maniadis, Tufano, and List 2014, 2017). Similarly, research efforts that would suffer from underpowered designs if conducted in a single locality can produce more defensible and insightful results when pooling real estate data across many sites.

Large-scale real estate datasets also create new analytical challenges. In the United States, such data are usually aggregated from public records provided by thousands of local data generators (county tax assessors, deed registries, and mapping departments). Analysts often find that large-scale public records data are provided in an only partially preprocessed state and require substantial cleaning to be usable for empirical analyses. For ZTRAX, Zillow explicitly cautions its users that "extensive exploring on your part is required due to the detailed, rich, and nuanced nature of the dataset" (Zillow Group 2021a). Researchers therefore face many data preparation choices that can affect findings but for which no published codebooks or best practice guidelines exist. If unreported, flexibility in data preparation adds to the range of "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011) that can leave studies vulnerable to researcher behavior that undermines the credibility of empirical findings (Christensen and Miguel 2018). Awareness of potential errors and biases, a full documentation of filtering choices, and a careful discussion of potential effects on research findings can reduce the influence of such "hidden" researcher decisions (Huntington-Klein et al. 2021), enable reviewers and editors to ask the right questions, and enhance the reproducibility, replicability, and generalizability of published work.

In this article, we catalog and discuss issues related to researcher decisions when working with real estate microdata. The article is the result of a group effort by academic researchers from 10 U.S. universities who have used ZTRAX for several large-scale property-level analyses and who share an interest in the accuracy, reliability, and reproducibility of empirical findings. Contributors to this article have used ZTRAX data to estimate the cost of land acquisitions for conservation purposes (Nolte 2020), the property value effects of national parks and historic sites (Zabel, Nolte, and Paterson 2024 [this issue]), the benefits of lake water quality (Mamun et al. 2023; Swedberg et al. 2024 [this issue]), the effects of water markets on agricultural land values (Chaudhry, Fairbanks, and Nolte 2024 [this issue]), the cost of hazardous chemical releases and the benefits of subsequent cleanups (Guignet and Nolte 2023; Guignet et al. 2023), the risk of flood damage to residential homes (Gourevitch et al. 2023), the effects of flood insurance policies (Hennighausen et al. 2023; Pollack et al. 2023), and the property value impacts of critical habitat designations under the U.S. Endangered Species Act (Mamun, Nelson, and Nolte 2024 [this issue]). Through this work, we have identified common problems of working with large-scale property data and experimented with potential solutions in the following areas:

- Identifying transaction prices reflecting fair market value.
- Geolocating transacted properties (land and buildings).
- Linking transactions to time-variant property characteristics.
- Identifying specific types of properties (e.g., single-family homes or vacant lands).
- Dealing with missing or mismeasured data for standard housing attributes.

After a brief introduction to data types and sources, we discuss each issue, establish its relevance, and consider a range of potential solutions for each. We then conduct a literature review of recent peer-reviewed ZTRAX-based analyses to examine the extent to which researchers disclosed their decisions on each issue. Last, we explore the extent to which findings can be affected by analysts' choices using an illustrative hedonic analysis with different data preparation specifications to estimate the effect of flood zone location on property prices. Alongside this article, we publish a set of digital resources (deed interpretations, filtering tables, source code) to document our choices and help other analysts implement, scrutinize, and improve data preparation procedures and assumptions.

While ZTRAX forms the basis of our analyses, the issues and solutions we discuss generalize to other real estate microdata sources. We see this article as a starting point for the development of best data practices guidelines for large-scale property-based analyses in the United States. Our propositions should not be interpreted as an attempt to develop universal standards for all cases, as many decisions will remain specific to research questions, location, and dataset. However, by helping researchers, reviewers, and editors develop an awareness for the potential consequences of data preparation choices on research results, we hope this article will encourage a broader application of steps that can increase the credibility and transparency of empirical findings, such as a more consistent documentation of data processing choices to facilitate reproduction and replication and a consideration of a broader range of errors and robustness checks in analyses and reviews.

2. Data Types and Sources

Real estate databases used in hedonic analyses often use at least one of three distinct types of public records data. In the Unites States, these public records are produced by different branches of local government (e.g., county, town, or registry district) and serve different purposes:

- Assessment data refer to tabular data compiled by local tax assessors for assessing and collecting property taxes. Because tax assessors are tasked with maintaining a complete account of the taxable value of all properties in their jurisdiction, assessment data can usually be expected to contain a complete or nearly complete list of properties in a county or town. The set of variables collected for each property varies across geographies but commonly includes property identifiers, such as assessor parcel numbers (APNs) and addresses; assessed or appraised values, sometimes provided separately for land and buildings; building characteristics, such as age, size, number of stories, bathrooms, bedrooms, and other features (pool, garage); land characteristics, such as lot size, land use category, and other features (e.g., lake frontage, views); and owner identifiers, such as names, addresses, and tax account numbers.

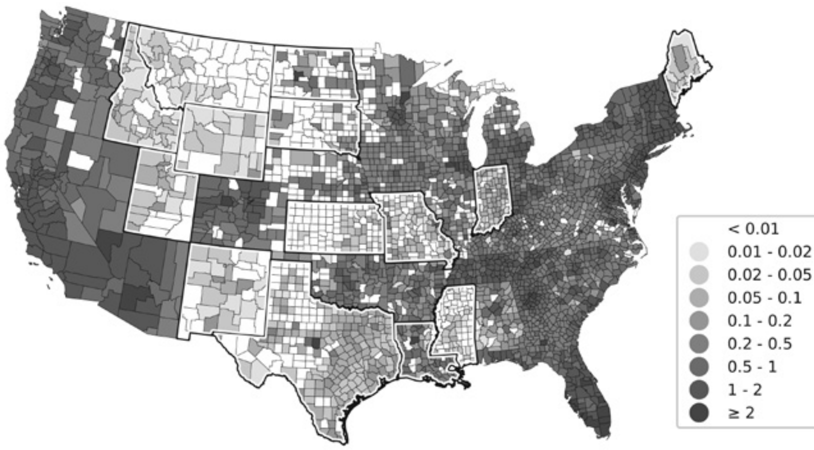
- Transaction data refer to tabular data of property transaction records, including deeds, mortgages, and foreclosures. Transaction data are generated only if and when property ownership changes and can include repeat sales (i.e., multiple transactions of the same property). Again, the set of available variables varies across geographies but often includes transaction price, date, document type (e.g., deed type), owner and seller names, property identifiers, and other information of interest (e.g., flags for intrafamily, arm's-length, or partial-interest transfers, information on mortgages and loans). Property identifiers can then be used to link assessment and transaction data.
- Parcel boundary data refer to geolocated polygons of parcel boundaries (i.e., vector data). Digital parcel maps now exist in all but a few U.S. counties. Parcel boundary data usually come with an attribute table that includes variables for each property, including parcel identifiers (APNs, addresses) as well as different subsets of attributes joined from assessment data, such as owner names and assessed values. Parcels and properties are not always identical: a single property can have multiple parcels (e.g., a large ranch), and a single parcel can include multiple properties (e.g., multifamily homes).

Our exemplar data, ZTRAX, contain assessment and transaction data but no parcel boundary data. Many of the insights we share have been obtained by comparing ZTRAX records with supplementary datasets, such as parcel boundaries, as well as satellite-derived building footprints and land cover classes. The authors affiliated with Boston University linked most ZTRAX records for the contiguous United States (CONUS) to parcel boundary data using text-based parcel identifiers and conversion algorithms developed as part of the Private-Land Conservation Evidence System (PLACES) (Nolte 2020) and described in a later section on geolocation. Because we use licensed parcel boundary data from third-party providers for approximately two-thirds of U.S. counties, we are not allowed to publicly share the full parcel-level dataset underlying our claims, such as corrected parcel and building coordinates. However, with the methodological descriptions we offer here and access to similar parcel boundary data, a computationally versed reader should be able to reproduce our findings for their study region and implement the proposed corrections and filters. Unless otherwise stated, all maps and statistics in this article are derived from the ZTRAX database version made available in October 2019 (downloaded on February 3, 2020) and limited to CONUS.

According to Zillow, ZTRAX is sourced "from a major large third-party provider and through an internal initiative we call County Direct" (Zillow Group 2021a). Geographic coverage of transaction data (>2,750 counties, >400 million transactions) is smaller than that of assessment data (>3,100 counties, >150 million properties). The dataset also contains an archive of historical assessment that allows tracking changes at the property level; its temporal coverage extends to the early 2000s in most states but varies across geographies (Appendix Figure A1).

ZTRAX exhibits substantial geographic heterogeneity in the availability of transaction price information, the dependent variable in most property-focused revealed-preference studies. The availability of price information is strongly shaped by the extent to which U.S. states require disclosure of sales prices (Figure 1). Lists of nondisclosure states commonly include Alaska, Idaho, Kansas, Louisiana, Mississippi, Montana, New Mexico, North Dakota, Texas, Utah, and Wyoming (Wentland et al. 2020). We also find sales price data to be rare in Indiana, Maine, Missouri, and South Dakota, and in a large share of counties in several other states (e.g., Alabama, Nebraska, Michigan, and Minnesota). Where the density of sale price observations is scarce, some transactions might still contain price data, but these are rarely representative (e.g., they might be associated with foreclosures or public sales) and therefore warrant greater

Figure 1
Density of Sales Transaction Data in ZTRAX
arms-length transaction prices / # parcels in county



Note: Density is computed as the county-level ratio of the count of nonduplicate arm's-length transaction records with prices greater than $1,000 and the count of parcels in digital parcel maps. Black-and-white boundaries show nondisclosure states or states with unusually scarce price data. The period spanned by ZTRAX transaction data varies across counties (Appendix Figure A2), which contributes to the observed differences in sales densities.

scrutiny. States and counties also vary in the length of the time for which transaction data are consistently available. Most counties contain transaction data for the first two decades of the twenty-first century. Three or more decades of data are available for the Northeast (Connecticut, Maryland, Massachusetts, New Jersey, New York, Rhode Island); much of California, Florida, Tennessee, and Ohio; and urban centers across the country (Appendix Figure A2).

3. Data Preparation Challenges

Identifying Transaction Prices Reflecting Fair Market Value

Real estate appraisals, hedonic pricing methods, risk assessments, and other property analyses often rely on the assumption that transaction prices are indicative of the fair market value (FMV) of the transacted property. FMV is the price at which a property would change hands between a willing buyer and a willing seller in a competitive market, neither party being under any compulsion to buy or sell. Public transaction data often include prices of transactions that do not fulfill these conditions, such as transactions between family members, transactions under distress (e.g., foreclosures), transactions below market value from public actors (e.g., by a targeted sale to veterans), or prices referring to monetary amounts other than the full property value (e.g., loans, mortgages, partial interests). To avoid biases that can affect subsequent conclusions, researchers must be able to identify and isolate FMV transactions.

Guidelines on how to identify FMV transactions in public records data are limited. Transaction records often include fields that can be used for developing filters, such as document types, seller and buyer names, or arm's-length or intrafamily flags. However, the interpretation of document types often requires domain expertise on the legal meaning and usage of different types of contract documents and how that usage varies across jurisdictions. Ancillary flags developed to address these concerns are often of undocumented provenance and incomplete. For instance, the transaction data in ZTRAX contain an intrafamily transfer flag that identifies transactions between family members using an undocumented algorithm. Based on a comparison of buyer and seller names, we estimate that this flag potentially misses 15.2 million intrafamily transactions (6.1% of deed records). Zillow confirms that their own cleaning procedure includes an internal text-matching algorithm (Zillow Group 2021a) that is not publicly documented.

We propose filters for nine variables available in ZTRAX, and likely in similar real estate databases, to identify transactions whose prices are more likely to reflect FMV (Appendix A1). Our approach distinguishes between transactions where the reported sales price reflects FMV with "high," "medium," and "low" confidence. After applying each filter to its respective variable, the individual filters can be combined (e.g., by only retaining transactions that obtained a "high confidence" value across all filters).

A description of filters and filtered categories is provided in Appendix B. Six of the nine listed filters are straightforward exclusions based on discrete values (data types, document types, loan types, price types, intrafamily flags) or cutoffs (for token prices). Three warrant elaboration:

- Name similarity: To enhance the identification of intrafamily transfers, we compute a similarity index between buyer and seller names. For each transaction, our algorithm computes the percentage of identical words appearing in both fields, weighing each word by the inverse of the square root of its relative frequency in the same county. This inverse frequency weighting reduces the probability that name similarity is erroneously established from frequently occurring words (e.g., John or Michael). We omit words with two letters or fewer and remove frequent generic words (e.g., bank, LCC). Transactions for which buyer or seller names returned a similarity index of 66% or more are flagged as "low confidence" FMV sales. We note that this threshold should be treated as a rule of thumb, as we do not have a way to validate predictions based on different cutoffs to determine which minimizes classification error. Using this method, we estimate that a naive approach to identifying intrafamily transactions from document types and the intrafamily transfer flag underestimates the actual extent of intrafamily transfers. Of 245 million deed records, ZTRAX flags 23.0% with its intrafamily transfer flags. We estimate the actual number to be 29.1%, a difference of 15.2 million transactions. We provide Python code to reproduce this similarity index with this article (Appendix E).
- Document types: Developing filters for the 161 different document types is not trivial because they can have different meanings and usages in different states. For instance, warranty deeds are the most frequent source of FMV transactions in most states, but grant deeds are more frequent in California, Nevada, and Vermont. Quitclaim deeds rarely contain sales price information in most states, except Massachusetts and Vermont where prices are frequently provided (Appendix Figure A3). Because ZTRAX contains 3,837 unique combinations of states and document types, an in-depth assessment of each combination is not feasible. Our filters therefore combine a hierarchical exclusion filter with a data-driven follow-up. First, with the help of a land use attorney, we make deterministic choices for the most frequent unambiguous document codes, including flags for foreclosures, intrafamily transfers, loans, and cancellations. An explanation of the meanings of the most frequent deeds is provided in Appendix A1. For the remaining, nonexcluded codes, we base our filters on two ancillary statistics computed for each combination of states and document codes: (1) the percentage of transactions with a price greater than \$1,000, and (2) the percentage of nonfamily transactions (see first bullet point). We flag state-code combinations where at least one of these statistics was found to be less than 10% or less than 33% as "low" and "medium" confidence, respectively.
- Public buyers and sellers: We identify public parties in a transaction from their respective name fields using text pattern matching ("regular expressions"; see Friedl 2006). There are a broad range of public organizations in the United States, and their names can be spelled in diverse ways in and across county registries. In states and regions where price data are scarce (e.g., Arkansas, Indiana, Texas), a comparatively large share of transaction data can

come from public sources of sales records, which can bias estimates of property value downward. Idiosyncrasies are common: in Arkansas, for instance, commissioners of state lands were identified in records by their personal names, not by their positions. Our approach is best described as a hybrid of top-down (names of preidentified agencies) and bottom-up (spelling learned from the data) identification of string patterns of public organizations. Owing to the absence of an external validation dataset for testing, we do not claim that the set of expressions is comprehensive. Instead, we share a machine-readable table of text patterns with this article (Appendix C) and will post feedback we receive at http://placeslab.org/hedonic-data-practices.

Geolocating Transacted Properties: Land and Buildings

Property analyses that study localized spatial phenomena often need accurate information of the location of land, buildings, or both. For example, many hedonic property value analyses infer landowners' preferences for environmental characteristics from spatial associations between the prices of transacted properties and environmental variables of interest. To do so, analysts must geolocate transacted properties and then computationally derive variables of interest from spatial data of environmental attributes. Depending on the application, demands on spatial precision can be high. For instance, Netusil, Moeltner, and Jarrad (2019) show that estimates of the impact of floodplain location on property values can be very sensitive to the choice of parcel boundaries versus buildings footprints as the spatial reference. Many analyses benefit from incorporating information on the characteristics of the land under each parcel. For example, estimates of the impact of development restrictions (and the cost of conservation easements) must account for a parcel's potential for future development, which is affected by its terrain, wetland presence, flood risk, existing land cover, and so on. Point locations are usually insufficient for the computation of such area-based proxies; instead, a geolocated polygon of the parcel boundary is required.

Many large-scale property analyses rely on point locations (latitude and longitude) found in assessment or transaction data to geolocate properties: for instance, 85% of peer-reviewed hedonic analyses using ZTRAX made this choice (see our subsequent literature review). However, the provenance and meaning of geographic coordinates are often unknown and poorly documented. Zillow describes its coordinates as "enhanced Tiger coordinates" (Zillow Group 2021a) and "populated by GEO-Coder," which might refer to the U.S. Census Geocoder (U.S. Census Bureau 2021). Using these coordinates without careful attention to the coordinate system, duplicates, missing data, and building locations can lead to misleading or unrepresentative findings. We identified six issues that are particularly worthy of attention. Appendix Figure A4 illustrates these issues in two U.S. counties (Middlesex in Massachusetts and Lane in Oregon):

- Latitude and longitude coordinates can be derived using various geodetic datums (e.g., NAD27, NAD83, WGS84). However, information about the geodetic datum used to derive coordinates is not always provided in the metadata of real estate databases, and in ZTRAX, it is entirely missing. Comparisons with georeferenced parcel boundary data suggest that until early 2020, the predominant datum of ZTRAX varied by county in a nonpredictable pattern (Appendix Figure A5). In more recent versions (October 2021), most counties use the more recent WGS84 datum, but exceptions remain (e.g., in New England). We also found counties using multiple geodetic datums for neighboring properties. Not correcting for these issues can lead to systematic geolocation errors that vary in magnitude across geographies: they will generally be higher on the coasts (~100 m in California, ~40 m in New England) but are largely negligible in the Midwest (e.g., Indiana and Michigan).
- Some coordinates seem to be derived from zip code area centroids instead of parcel locations or street addresses, without being flagged as such. Anecdotal evidence from visual inspection suggests that this issue is particularly common for recent

subdivisions and properties without addresses. Using these coordinates can lead to geolocation errors of greater than 1 km.

- Many records are missing latitude and longitude data (Appendix Figure A6). Missing data are often associated with particular types of parcels (e.g., vacant parcels, rural parcels, records without addresses). Excluding them from an analysis where such parcels would otherwise be included will result in nonrandom selection into the sample.
- Some counties appear to base their coordinates on parcel centroids, whereas others seem to refer to building locations. Distances between building footprints and parcel centroids vary across the country (Appendix Figure A7); mean distances of more than100 m are common in rural settings with large parcels. Uhl et al. (2021) compared ZTRAX locations with remote sensing–derived building footprint data (Microsoft 2018) and find positional accuracy decreases as one moves from urban to rural settings. Analyses that assess the impact of spatially precise policies (e.g., official floodplains) are thus subject to errors of possibly large magnitudes (Netusil, Moeltner, and Jarrad 2019).
- Most counties contain at least some incorrect, nonduplicate parcel locations (Appendix Figure A8). Possible observed reasons include coordinates being based on owner's mailing addresses (instead of property location addresses) and subdivisions of parcels.
- Point locations can change between updates. For instance, in a comparison of Rhode Island property locations between versions of ZTRAX downloaded in 2017 and 2019, we found zip code area centroid placeholders replaced with street address geolocations (Appendix Figure A9). For many Rhode Island properties, we found minor shifts in point locations that were multidirectional and thus not simply attributable to changes in projection (Appendix Figure A9). Although such changes appear to reflect improvements in geolocation over time, they also highlight the need to exercise caution when geolocating assessment data that span multiple time periods.

There are several options to improve the geolocation of assessment and transaction data. Because geocoordinates appear to have improved in recent years, we recommend starting with the most recent available database. Analysts can choose among the following options as a function of their resource constraints (data access and time available for data inspection and cleaning) and the anticipated sensitivity of findings to geolocation errors:

- Quick fixes: Analysts without access to digital parcel maps or without the geoprocessing skills to link assessment and transaction records to parcels and buildings can enhance the reliability of their findings with two fixes: ensure that the correct geodetic datum is used to spatially locate coordinates (Appendix Figure A5, Appendix D) and drop entries with duplicate coordinates, especially if these coordinates are zip code area centroids and if dropped records have either no or unique street addresses. Appendix Figure A6 shows the prevalence of missing and nonempty duplicate coordinates in ZTRAX assessment data and thus of anticipated reductions in sample size and county coverage when removing these entries. Appendix Figure A8 shows how much geolocation error remains in each county after implementing these two fixes. In most counties, the median geolocation error drops to less than 1 m, suggesting that most parcels are correctly located. However, mean errors can be large (often >500m), indicating that a share of parcels will remain incorrectly located, sometimes to a large degree.
- Crop to county and zip code boundaries: If county identifiers and zip codes are provided, they can be used to remove coordinates that fall outside the corresponding spatial boundary. Official boundaries of counties and zip codes are available through the Integrated Public Use Microdata Series National Historical Geographic Information System (IPUMS-NHGIS) (Manson et al. 2018). We do not recommend cropping to smaller census units (e.g., census tract or block boundaries): in the case of ZTRAX, identifiers for these units appear to have

been derived directly from the geocoordinates through spatial joins.

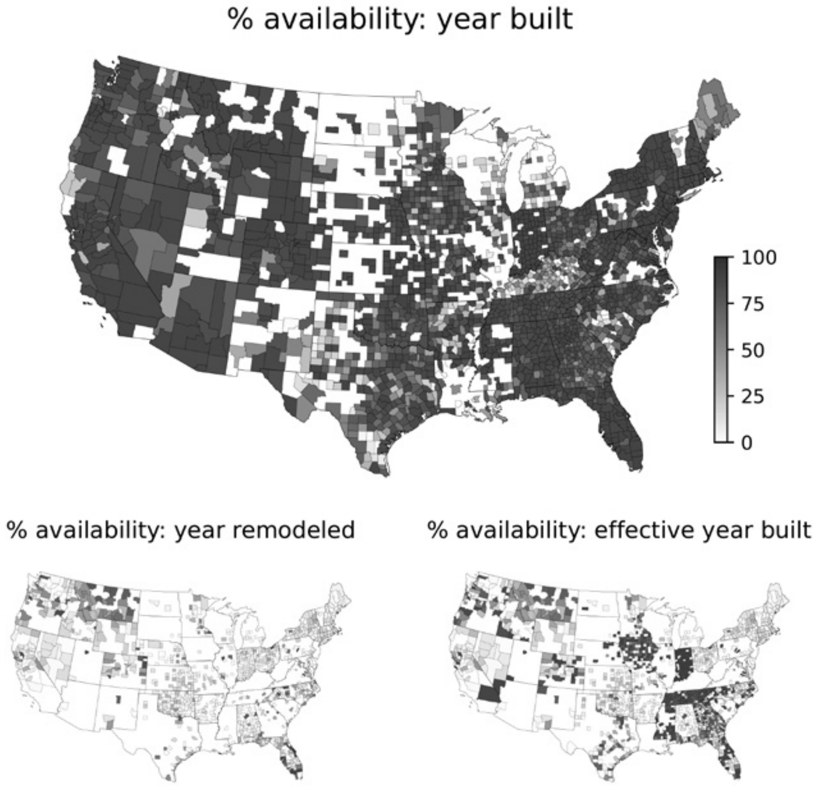
- Geocode addresses: If records with missing geocoordinates contain addresses (street, city, and zip code) (Appendix Figure A10), analysts can improve the completeness of geolocations by means of additional geocoding (e.g., by using the U.S. Census Geocoder or the GeoCoder API, at https://geocoder.readthedocs.io). This approach remains untested and might be vulnerable to the same issues we observe for coordinates in assessment data but will likely be able to take advantage of recent updates to geolocation databases.
- Linking assessment data to parcel boundary data: Digital parcel maps now exist in at least 3,073 (97.8%) of counties in CONUS. In most cases, parcel boundaries can be uniquely and reliably linked to assessment data using APNs or unique taxpayer account numbers. This approach tends to lead to a more reliable and complete geolocation of assessment data than the previous fixes. It also allows analysts to derive important indicators of property value from the parcel boundary data (e.g., building footprint size, road access, lake frontage, wetland coverage, forest stocks). However, establishing this linkage is complicated by idiosyncratic differences in the syntaxes of APNs, which vary between assessor and parcel boundary datasets and geographically between neighboring counties and towns. We recommend a four-step approach that consists of (1) developing text pattern descriptors (regular expressions; see Friedl 2006) to identify the prevalent APN syntax in a given county or town and to extract the identifying text fragments (e.g., numbers or letters without spaceholders); (2) reformatting and recombining the extracted text fragments to create a new parcel identifier with the same syntax for both assessment and parcel polygon data; (3) iterating over the two previous steps until the new parcel identifier produces the largest number of uniquely matched records across assessment and parcel boundary data; and (4) double-checking that this linkage results in relatively small spatial distances between geographic locations provided in the assessment data and matched parcel boundaries. The last step helps identify erroneous linkages when the syntaxes of two identifier columns are apparently similar but refer to different concepts (e.g., both APNs and tax account numbers might use numeric identifiers, but one refers to parcels and the other refers to persons). Using this approach, analysts will be able to link most parcel boundaries to assessment data in most counties (Appendix Figure A11).
- Identifying building locations: After linking assessment records to parcel boundaries, analysts can use spatial data on building footprints to identify the precise location of buildings within a parcel. A nationwide open-source dataset of 130 million building footprint polygons was made available by Microsoft (2018) and has been updated multiple times since its initial release. Derived from high-resolution satellite imagery with a documented machine learning algorithm, these data are, to our knowledge, the most consistent open-source indicator of building presence for the United States currently available free of charge. Some downsides remain: dates of observation are not always provided and can be more than a decade old in some instances. We also observe an underreporting of buildings under tree cover.

Linking Transactions to Time-Varying Property Characteristics

Analysts often need to establish a reliable link between transaction prices and the characteristics of the property at the time of sale. Assessment data are an important source for these attributes, often reporting building square footage, lot size, architectural style, counts of units, rooms, bedrooms, bathrooms, and the presence of other features (e.g., garage, pool). Parcel boundary data can provide further information on lot size, building location, land cover, and access to roads, water bodies, or open space. These characteristics can change over time as buildings are built, remodeled, or destroyed and as boundaries are redrawn following subdivisions or mergers. Analysts usually want to be certain that characteristics observed in assessment data and

Figure 2

County-Level Availability of Data on Years of Change to Buildings in Residential Assessment Records in ZTRAX: *top*, Year Built; *bottom left*, Year Remodeled; *bottom right*, Effective Year Built.



Note: The "effective year built" is of unknown provenance; it could be a hybrid of "built" and "remodeled" year but might include other adjustments.

parcel boundary data are the same as those observed at the time of sale

Assessment and parcel boundary data provide only cross-sectional snapshots of property conditions at a single point in time, typically a recent one. Dataset versions for multiple years sometimes exist, and some providers of parcel boundary data offer archives of historical data. However, synthesizing datasets from multiple time periods substantially increases data volumes and time cost for an analysis, often with uncertain benefits. Furthermore, not all regions have historical data.

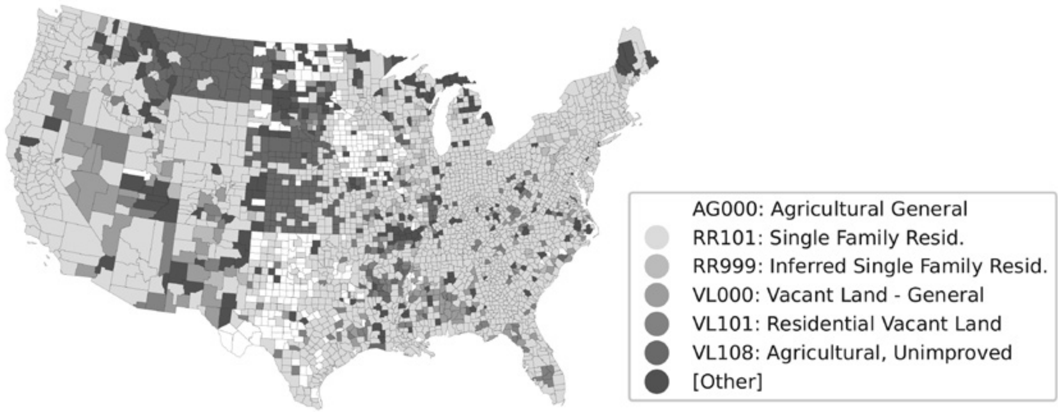
Analysts often consider alternative strategies to exclude transactions whose observed characteristics might not reflect those at the time of sale. Unobserved renovations between sales are particularly problematic for repeat sales analyses, often considered a gold standard for evaluating property value changes (Bishop et al. 2020; Banzhaf 2021). The availability of data on the time a building was built or remodeled varies across the United States (Figure 2; see also Leyk et al. 2020). We also observed building year values for identical properties differ between historical and current versions of assessment data in bidirectional and idiosyncratic ways that cannot be explained by new constructions alone (Appendix Figure A12) but indicate that building year data are collected, updated, and interpreted in different ways across space and time. Filtering choices that increase the confidence in the quality of data over time (e.g., dropping counties, dropping sales, ignoring the issue) will likely affect the geographic coverage of findings (e.g., dropping observations in Vermont or Wisconsin). Satellite-based land cover change observations offer the potential of an independent detection of changes but come with their own set of challenges, such as classification errors, insufficient resolution, or limited temporal coverage.

In our analyses, we consider the following solution options. Their relative utility to the analyst will depend on the application and study area. For instance, analyses of the temporal dynamics of urban growth will likely need to apply more rigorous standards than analyses of the effects of changes to nearby amenities in a stable urban core.

- Identify and account for sales with misrepresented characteristics based on years of building updates: Assessment data often contain information on the year the building was first constructed and in a minority of counties the year of the last remodel (Figure 2). Where both variables are available, analysts can identify transactions of properties that have been developed or remodeled since the sale. Based on available data, such sales make up 2%–10% of the sample in most counties (Appendix Figure A13). There are two possible approaches to account for these transactions: (1) running the analysis after excluding such observations, and (2) controlling for an indicator of such transactions and interacting it with all hedonic variables. These approaches provide useful robustness checks that analysts can use to gauge the importance of potentially misrepresented variables in the context of their analysis. Counties that provide data on building year allow for the exclusion of new developments, but analysts will need to consider the probability of remodeling and potential biases as part of their estimation procedure. In counties where neither type of data are available, the analyst must also consider the extent to which new unobserved buildings might affect their results. Finally, pending a more in-depth understanding of the reasons behind idiosyncratic changes of building year data over time (Appendix Figure A12), analysts might consult with local tax assessors about the reasons for such changes or conduct sensitivity checks that incorporate building year data from different database versions (including the database history) or external data sources, such as historical maps or aerial imagery.
- Exclude sales based on remote observations: In the absence of consistent building year indicators, we considered leveraging public, satellite-based indicators of land cover change of increasing spatial-temporal resolutions and extents. Unfortunately, most nationwide historical estimates before 2013 will likely be based on products derived from medium-resolution imagery (Landsat) that are not always reliable (Brown et al. 2020) and often miss low-density development in rural, forested areas (Olofsson et al. 2016). Using the most recent public release of LCMAP, a product developed by the U.S. Geological Survey that tracks annual change to land cover between 1985 and 2017, we find low correspondence between building years in assessment data and remotely sensed transitions from undeveloped to developed land cover (Appendix Figure A14). Modern high-resolution satellites with more frequent temporal coverage (Sentinel-2, Planet Labs) will help improve observations of change. Because of the observed uncertainties associated with this approach, we currently recommend it as a robustness check only.
- Constrain the time horizon of the analysis: We expect the likelihood of unobserved changes to be higher the more time has passed since sales and the observation of property characteristics. Analysts can narrow the time horizon of the analysis by excluding sales outside a time window around the acquisition date of the property data. This likely reduces error but at the expense of a reduced sample size, a lesser ability to observe long-term trends, and lower explanatory power of analyses estimating effects of natural events or policy changes that happened further in the past.
- Use datasets from multiple time periods: Historical assessment and parcel boundary data can sometimes be obtained from data aggregators. We have not systematically assessed the availability and quality of such historical data across the country, but we anticipate that they vary geographically

Figure 3

Most Frequent Land Use Code across All Parcels Linked to ZTRAX Assessment Data in Each County



as a function of the time at which county offices digitize their records and data aggregators expand their geographic coverage (a process that is still ongoing).

Identifying Different Types of Properties

Analysts often need to be able to restrict the sample of transactions to specific types of properties, such as single-family homes, agricultural lands, or vacant lots. Hedonic valuation studies require the identification and delineation of a real estate market to satisfy underlying assumptions that identical properties will sell for the same price throughout that market (i.e., the "law of one price"; see Bishop et al. 2020). Similarly, efforts to estimate the value of undeveloped land (Nolte 2020) must be able to reliably identify and exclude parcels with buildings, as buildings often represent a large share of a property's value.

The availability, usage, and quality of variables used to identify properties in submarkets vary across geographies. For instance, ZTRAX contains a "property land use standard code" with a hierarchical classification scheme, but its provenance is undocumented, and the type of properties identified with a specific code varies, often across state boundaries (Figure 3). In most counties, single-family homes are identified as RR101, but in others, RR999 (inferred single-family), RR000 (general residential), or RR102 (rural residence) are also commonly used. Some counties have one code for all agricultural land (AG000), while others use VL108 (agricultural, unimproved) to identify similar lands or break down the agricultural land category into subcategories such as AG101 (farm) and AG109 (timberland/forest/trees).

Identifying vacant lands is particularly challenging. For instance, ZTRAX's indicator "number of buildings" misses most buildings in hundreds of counties (Appendix Figure A15). A substantial share of parcels with "vacant" land use codes have building footprints, and many parcels with "residential" land use codes have no building footprints (Appendix Figure A16). Alternative indicators for building presence can be derived from assessment data (e.g., property land use codes or the assessed value of buildings) or from remote sensing data linked to parcel boundaries (e.g., building footprints or developed land cover). However, no single indicator is unambiguously perfect for a nationwide analysis (Appendix Figure A15).

We recommend that analysts of assessment data exercise caution in developing their submarket filters by testing the robustness of their results to alternative plausible filtering conditions, and documenting their choice of filtering steps alongside published results.

- Single-family homes are likely best identified by combining several land use codes (in ZTRAX: RR000, RR101, RR102, RR999) with indicators confirming the presence of a building, such as a positive assessed or market value for buildings, square footage, gross building area, or the presence of a remotely observed building footprint on the parcel. We note that the presence of a building does not necessarily guarantee that the building is a single-family home. We recommend that analysts double-check whether, in their study area, land use codes are based on legal zoning or imply the presence of a building.

- Vacant parcels (parcels without any buildings) are most reliably identified through a combination of multiple variables. Analysts wishing to minimize the likelihood of an erroneous inclusion of buildings can exclude parcels without building footprints, without a land use code indicating the presence of a building, and without a positive value for improvements in either the tax assessors' valuation or their FMV estimates.
- The identification of agricultural parcels also benefits from combining multiple variables and can be enhanced through a judicious use of other data sources. Reducing omission and commission errors is particularly important for this category, as agricultural land markets are thin, with only a small fraction of agricultural land sold annually (Bigelow, Borchers, and Hubbs 2016), and small errors can lead to small sample sizes or bias. Agricultural parcels can be found under a range of land use codes; in ZTRAX, these include AG (agricultural), VL (vacant land), and RR (residential). Analysts filtering assessment and transaction data for agricultural sales might consider additional variables—such as lot size, location, and the relative size of building footprints—as indicators of agricultural properties. Attention to regional agricultural production and institutional detail is important. For instance, in the western United States, where irrigation is particularly important for agriculture, spatial layers for identifying irrigated cropland from governmental sources can help distinguish between irrigated and nonirrigated agricultural areas in sample selection and analysis.

Dealing with Missing or Mismeasured Data for Standard Housing Attributes

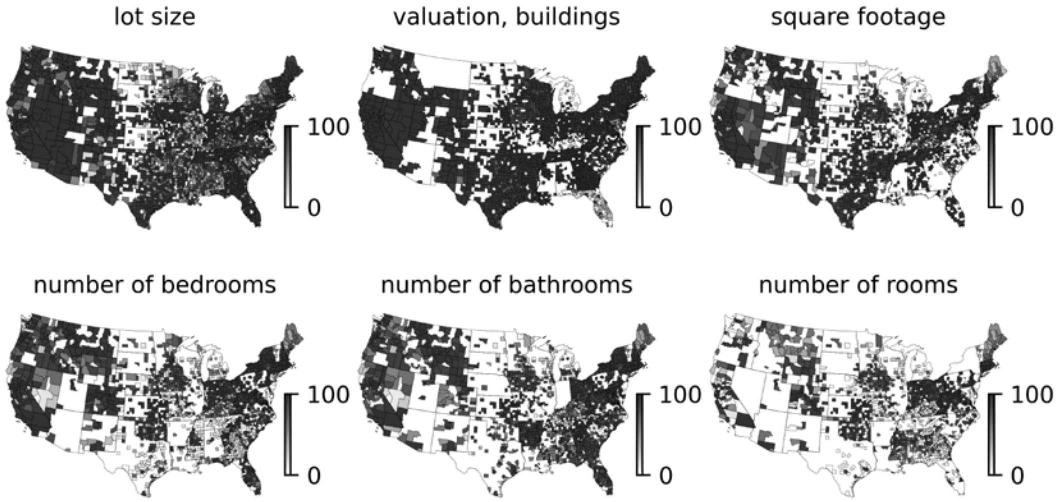
Hedonic analyses must distinguish effects of environmental attributes on property values from those of other confounding variables. Many analyses control for key characteristics of land and buildings in a regression framework and/or through matching techniques. This requires that these characteristics are reliably and consistently observed across the study region.

The availability of standard housing attributes in assessment data varies across counties, often clustered by state. For instance, across all residential property records in ZTRAX, data gaps exist for lot size (15.1% missing), building valuation (21.4%), square footage of living area (28.6%), number of bathrooms (33.1%), number of bedrooms (53.1%), and total number of rooms (60.1%) (Figure 4). Nonsensical zero values (e.g., zero rooms, zero living area) are not uncommon. When nonzero values are observed, they can refer to different units or measurement strategies, which are not necessarily explained (e.g., frontage feet vs. lot area, square footage of building footprint vs. square footage of all floors). For instance, despite the nearly complete availability of "lot size" data (Figure 4), summing the lot size of all parcels in a given county in Florida does not aggregate to the total area of the county (Clapp et al. 2018). Differences can occur across and within jurisdictions, presumably due to variability in practices between communities or individual assessors.

Missing data means that researchers face sample selection issues, whereas unreliable data is a measurement error issue. Both can involve trade-offs between empirical specification and geographic coverage. For instance, analysts who prefer to exclude records with missing data will likely have to work with a substantially constrained and geographically nonrepresentative sample. A pairwise completeness analysis for a subset of attributes (Appendix Figure A17) indicates considerable heterogeneity of attribute completeness in assessment data; for example, the joint analysis of building square footage and land use type would cover a sample of 68% of its almost 150

Figure 4

County-Level Availability (Percentage of Nonzero Values) of Standard Housing Indicators for Residential Parcels in ZTRAX Assessment Records



million property records. Analysts who maintain those observations and somehow account for missing attribute values in their empirical models must consider how their choice of methods might bias their estimators and affect the geographic extent of their analysis. Common examples of these methods include using only a subset of the available measures; using spatial or temporal fixed effects, the average housing characteristics in the location, or repeat sales models to proxy for unobserved quality; using dummy variables to control for missing observations; and interpolating missing values either from available indicators or based on out-of-sample data that contain new information (Moulton, Sanders, and Wentland 2018; Clarke and Freedman 2019; Gindelsky, Moulton, and Wentland 2019; Albouy, Christensen, and Sarmiento-Barbieri 2020; Fraenkel 2020). Analysts working with temporally varying characteristics from the historical assessment data need to be aware of data gaps resulting from subcounty level updating cycles of the underlying assessment data that lead to incompleteness patterns that vary across space and time (Appendix Figure A18a). Methods to mitigate the resulting bias may include spatial aggregation (Appendix Figure A18b) or record-level time series interpolation (Appendix Figure A18c).

A full assessment of the performance of different approaches to account for missing and mismeasured data in housing market models is beyond the scope of this study. Cameron and Trivedi (2005) offer a discussion of these issues and potential solutions. The key issue is understanding whether data errors are random or systematic. Determining how data errors vary spatially will help analysts account for these issues in their empirical specification. We recommend that, even in the presence of time constraints, analysts dedicate a significant amount of time to inspecting data, applying robustness checks, and fully documenting choices and findings.

Data inspection can range from simple "sanity checks" (e.g., verifying the plausibility of values with histograms, checking for unexpected clustering with maps) to more systematic testing, such as calculating correlations between data issues and the outcomes of interest, observable characteristics (e.g., jurisdiction, income, race), or matched external validation data (e.g., lot sizes from parcel boundaries, jurisdiction, income, race). The appropriate data inspection approach should be guided by the analyst's research question and design. For instance, the pattern of missing and mismeasured data that cause bias are likely to be different between cross-sectional and panel or difference-in-difference models (see the discussion in Zhang, Phaneuf, and Schaeffer 2022).

Analysts should also include a suite of robustness checks involving different plausible combinations of data filters and models and examine the sensitivity of findings to their choices.

Most important, we recommend that analysts fully document their sampling procedure, including choices of inclusion versus exclusion of observations and attributes based on data availability and reliability, and the implications of that choice on the geographic coverage of findings.

4. Choices Reported in the Literature

To what extent do current hedonic analyses of environmental attributes already acknowledge and address these challenges? To answer this question, we reviewed data filtering, processing, and modeling choices reported in the 27 peer-reviewed journal articles that used ZTRAX to value an environmental attribute using hedonic property value methods and were published by August 10, 2022 (Appendix Table A1). We identified this sample by searching for the term "ZTRAX" in Google Scholar and retaining from the resulting 320 records all studies that met these criteria. We checked whether each study reported undertaking any step from a list of 35 individual processing steps across the five previously discussed challenges (i.e., observed positives; see Figure 5).

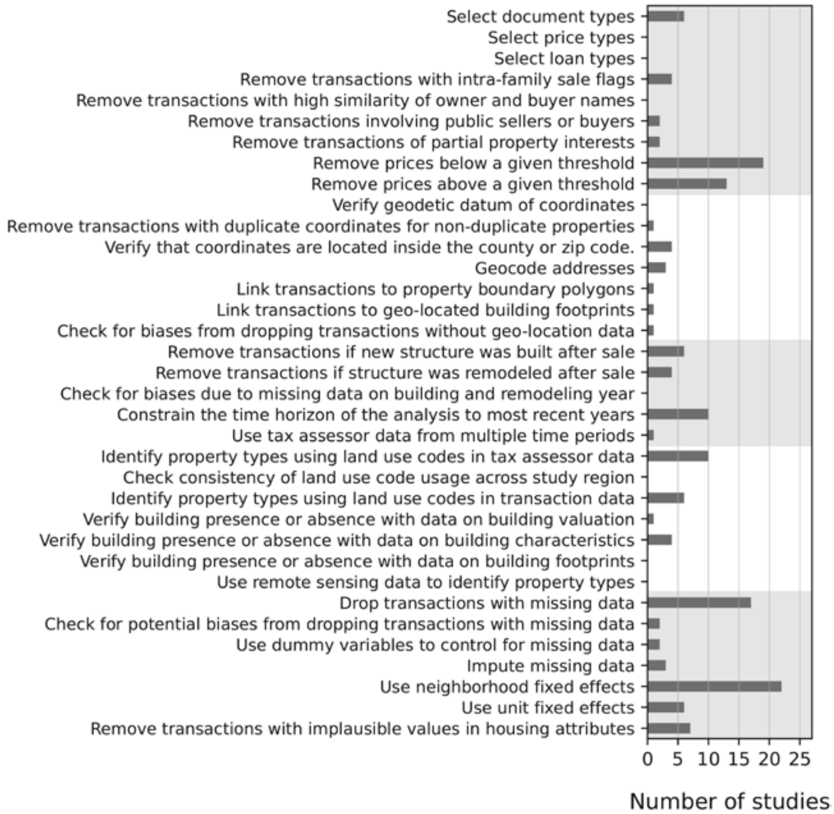
On average, we find that reviewed studies report only a small fraction of the proposed steps (average: 5.85 of 35 potential steps; range: 1–11) (Figure 5). Although small, this number does not by itself cast doubt on the validity of the findings of any study for at least three reasons: authors might have implemented a step without reporting it, not all steps are necessary in every analysis (e.g., missing data can be negligible in a given study region), and some steps are substitutes (e.g., filtering out implausible coordinates vs. linking records to parcel boundaries). However, our review also suggests that some peer-reviewed articles might have cut short the full reporting of relevant choices inherent in the analysis of large-scale property data, creating additional and likely unnecessary barriers to reproducibility and replication.

In terms of individual challenges, we find:

- Arm's-length sales filters: Upper and lower thresholds on sales prices are the most frequently reported type of filter. Thresholds vary widely (lower: \$1–\$100,000; upper: \$1–\$10 million or top 0.5%–5%). No study reports excluded sales based on a high similarity between buyer and seller names or based on price types or loan types.
- Geolocation: 23 studies (85%) use property coordinates to measure spatial relationships to the environmental attribute of interest, but only eight report taking any step to address potential geolocation issues, and none report verifying the geodesic datum of coordinates. Authors of three studies chose to ignore property coordinates provided in assessment or transaction data, geocoding street addresses instead.
- Time-varying characteristics: A few studies report to have excluded sales that occurred before a house was built ($n=6$) or remodeled ($n=4$). It is also common for studies to reduce the time horizon to a more recent time (e.g., 10 studies chose ≥2005).
- Property types: 26 of 27 studies focus on residential properties, predominantly single-family homes. However, only a fraction report how the sample was selected, and only four take any steps to examine omission errors (e.g., by not including the full set of building codes) or commission errors (e.g., by verifying that a building exists).
- Missing and mismeasured housing attributes: Most studies include some housing attributes in their hedonic regression ($n=25$), including living area ($n=19$), age ($n=19$), bedrooms ($n=18$), bathrooms ($n=18$), and lot size ($n=15$). The most common choice to deal with missing attributes is to drop observations with missing data ($n=17$), whereas only two studies add missing value indicators and recode missing values as zeros. Most analyses also

Figure 5

Number of Published Peer-Reviewed Studies Reporting Implementing a Given Data Processing Step from an Environmental Hedonic Analysis Based on ZTRAX Data



Note: The total number of studies is 27, and data are noted as of August 10, 2022.

include neighborhood fixed effects to control for unobserved attributes ($n=22$). The smallest spatial scale of these fixed effects varies across studies: zip codes are most frequent ($n=6$), followed by block group ($n=3$), tracts ($n=3$), and counties ($n=3$).

5. Sensitivity of Hedonic Coefficients to Data Preparation Choices: An Illustration

Are the results of hedonic studies sensitive to whether and how an analyst chooses to address the five challenges we outline? The answer depends on many factors specific to a study, including its objective, geographic scope, inferential strategy, and coefficients of interest. We therefore cannot address it comprehensively. Instead, we use an illustrative case study to explore whether data processing choices matter in at least one application of interest.

Our case study focuses on the property price effects of being located inside a special flood hazard area (SFHA, 100-year flood zone), as mapped by the Federal Emergency Management Agency (FEMA) (Bin and Kruse 2006; Bin, Kruse, and Landry 2008; Beltrán, Maddison, and Elliott 2018). Because we are interested in highlighting the importance of county-level variation in data availability and quality, we estimate this effect separately for each county within the CONUS.

In each county, we estimate the following log-linear regression:

$$\ln(price_{ijt}) = \alpha + \delta\, SFHA_i + X_i\beta + \mu_j + \tau_t + \varepsilon_{ijt}, \quad [1]$$

where $price_{ijt}$ is the sales price of property i in neighborhood j at time t. The indicator variable of interest is $SFHA_i$, which is one if the sold property was located inside the SFHA ("treated"), and zero if FEMA considered the property to be located outside the SFHA ("control"; unmapped areas are excluded). Therefore, δ is the coefficient of interest. X_i contains property-level characteristics, μ_j are neighborhood (spatial) fixed effects, τ_t are year-quarter fixed effects, and ε_{ijt} is an error term. Definitions of $SFHA_i$, X_i, and μ_j vary across our sensitivity checks:

- In our default model, $SFHA_i$ is based on Microsoft building footprints: it is one (treated) if the centroid of the largest building footprint on a given parcel is located inside the SFHA, and zero otherwise. We include only properties with one or two footprints. As sensitivity checks, we derive $SFHA_i$ from either parcel boundary centroids or assessment data coordinates (October 2021 version of ZTRAX); in both cases, we also assume that the analyst did not use any building footprint data to drop observations without observable buildings or with more than two building footprints.
- X_i includes lakefront and riverfront indicators to control for the amenity of water access. We derive both from spatial proximity of parcel boundaries and waterbody polygons from the National Hydrography Dataset (U.S. Geological Survey 2017). In our default model, X_i also includes bedroom and bathroom dummies (i.e., dummy variables for 1, 2, … 10 bedrooms and 0.5, 1, … 10 bathrooms). We vary X_i across two sensitivity checks: one that drops bedroom and bathroom dummies and one that adds total living area as a supplementary control. Our default run keeps sales of properties with missing bedroom and bathroom data, and adds separate dummies for missing values, zero values, and values above 10 to each (i.e., 2 × 3 dummy variables). A sensitivity check drops observations with missing bedroom or bathroom data or zero values in either field.
- In our default model, μ_j stands for census tract fixed effects. As sensitivity checks, we also switch to zip code (coarser), block group (finer), and no spatial fixed effects. All spatial units are derived from 2016 NHGIS data 9 (Manson et al. 2018).

Including sales in the regression is subject to multiple filters. In all cases, we select sales of single-family homes that occurred in or after 2000 or after the most recent update to the SFHA in their county (whichever is later) and before October 1, 2021. Owing to well-documented difficulties in separating the negative price effects of coastal flood risk from the positive price effects of coastal amenities (Beltrán et al. 2018; Johnston and Moeltner 2019), we exclude sales located within 2.5 km of an ocean coast. In addition:

- Our default model only includes sales that pass all our high confidence arm's-length filters. In two sensitivity checks, we also include medium and low confidence sales. In addition, we test the effects of using only a lower price threshold (≥ $1,001), the most frequently reported filter in our literature review.
- Our default model drops sales of properties whose buildings were known to be built in the same year or after the sales transaction occurred. A sensitivity check drops this filter.
- Our default model uses several document codes to identify single-family homes (RR000, RR101, RR102, and RR999). A sensitivity check uses only the simple single-family flag (RR101).

To illustrate the joint importance of multiple decisions, we derive a "current literature" scenario with a set of changes to our default model that we consider representative of steps reported in the existing ZTRAX-based literature: (1) the only arm's-length filter used is a lower price cutoff ($1,001), (2) treatment identification ($SFHA_i$) relies on property coordinates in the assessment data, (3) no building footprint data are used to select the sample, (4) sales with empty or extreme values for bathrooms and bedrooms are dropped, and (5) sales with new buildings since the transaction are kept.

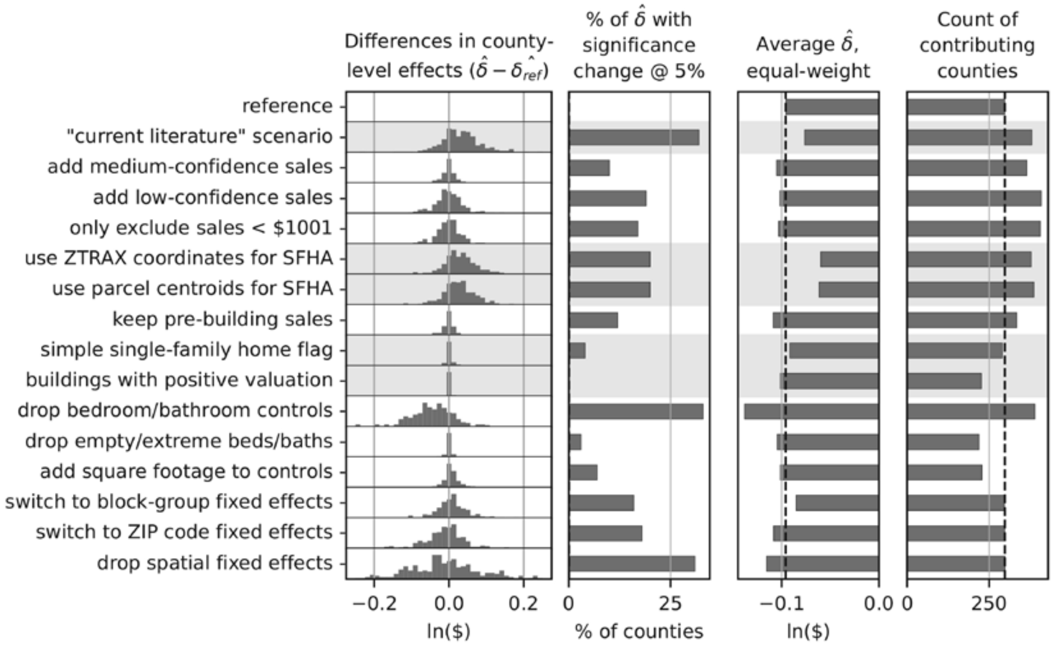
After fitting each model at the county level, we keep results from all county-level models with a minimum of statistical support, which we define here as being estimated on a sample containing at least 100 identifying treated and 100 identifying control sales. With "identifying," we mean that we count only sales belonging to categories (lakefront, waterfront, year quarter, bedroom count, bathroom count, and spatial fixed effect, if applicable) that exhibit treatment heterogeneity (i.e., that contain both treated and control sales). While our default model retains results from 297 counties, this count can range from 219 counties when dropping sales with empty or extreme bedroom or bathroom counts (variables for which data are often missing, see Figure 4) to 409 counties when adding in transactions that had not passed our arm's-length filters.

We compute all county-level differences in the estimated coefficient of interest ($\hat{\delta}$) between our default model ($\widehat{\delta_{\text{ref}}}$) and each alternative model specification with the same minimum of statistical support. To assess the nationwide magnitude of the effect of processing choices on the estimated SFHA discounts, we also report the percentage of counties in which a county-level study would have led to different conclusions regarding the statistical significance of the SFHA discount (at $\alpha = 0.05$) as well as the averages of county-level estimates of $\hat{\delta}$, weighted equally by county ($\hat{\delta}_{\text{avg}}$).

Our results demonstrate that each processing choice has discernible effects on the magnitude and significance of $\hat{\delta}$ or the geographic coverage in our application (Figure 6). Importantly, we find that the effects of choices that are rarely reported in the peer-reviewed hedonic literature—such as the choice of arm's-length filters, geolocation precision, or the removal of prebuilding sales—can be of similar magnitude as the effects of choices that are more commonly reported, such as dropping missing data observations or using different spatial fixed effects. For example, 17% of county-level estimates cross the statistical significance threshold of $\alpha = 5\%$ (i.e., switch significance in either direction) when using only a minimum price filter for arm's-length sales, 20% when switching to property coordinates in assessment data instead of building footprints to assign SFHA treatment, and 12% when not removing prebuilding sales. These counts are of a similar order of magnitude as those observed when switching to zip code fixed effects (18% of county effects switch significance) or block group fixed effects (16%), and substantially larger than the consequences of dropping missing data observations (3%). Similarly, effects on the (county-weighted) magnitude of the SFHA discount can be large. Using only a minimum price filter for arm's-length sales increases the absolute value of the discount, ($\left|\hat{\delta}_{\text{avg}}\right|$) by 8%, switching to property coordinates in the assessment data reduces it by 36%, and not removing prebuilding sales increases it by 13%. Meanwhile, switching to block group fixed effects decreases its absolute value by 11%, switching to zip code fixed effects increases it by 13%, and the dropping of missing data observations increases it by 9%.

Common strategies to strengthen the robustness of empirical estimates do not fully remove the observed sensitivity of $\hat{\delta}$ to data processing choices. For instance, if we reduce our sample to the counties with particularly strong statistical support (≥ 500 identifying treatment and control sales, $n = 73$), our estimates remain sensitive to most specifications (Appendix Figure A19). Similarly, if we only retain counties whose estimates of $\hat{\delta}$ are robust to variations in fixed effects (defined here as $\left|\hat{\delta} - \widehat{\delta_{\text{ref}}}\right| < 0.02$ when switching to block group and zip code fixed effects, $n = 69$), many estimates remain sensitive to data processing choices about arm's-length filters and geolocation (Appendix Figure A20). A third potential strategy to enhance the robustness of estimates is to pool the data across larger geographic units, such as states. We repeat our full analysis at the state level for all states that contain at least 500 identifying treatment and control sales ($n = 32$). Although we find that this approach greatly reduces the number of changes to the statistical significance of state-level estimates when changing spatial fixed effects (only one estimate switches, $\alpha = 0.05$), several results remain affected by choices on

Figure 6
Effects of Variations in Data Filtering, Processing, and Modeling Choices on Estimated Flood Zone Discounts



Note: Dashed lines mark the value of the reference model.

arm's-length filters and geolocation (Appendix Figure A21).

A full assessment of the mechanisms behind the observed sensitivities is beyond the scope of this article, but we search for potential reasons by closely examining the data for a small set of counties whose estimates of $\hat{\delta}$ are particularly sensitive to data processing choices despite strong statistical support. In Montgomery (Pennsylvania), Polk (Florida), and Fulton (Georgia) Counties, we find that many property boundary polygons are defined such that their centroids fall inside the SFHA, although their buildings are located outside (Appendix Figure A22). This leads to an underestimation of the absolute value of $\hat{\delta}$ when using parcel centroids to allocate treatment, a finding in line with Netusil, Moeltner, and Jarrad (2019). In Rock Island County, Illinois, we find that relying on coordinates in assessor data to identify SFHA location overestimated the absolute value of $\hat{\delta}$ because the county's assessor data have numerous missing and erroneous coordinates (zip code centroids, owner mailing addresses) whose flood zone location was associated with sales prices. In flood zones, erroneously located sales had lower average prices than correctly located ones, while outside flood zones, erroneously located properties had higher average prices than correctly located ones. Finally, in Bucks County, Pennsylvania, we find that using lower price cutoffs as the only arm's-length filter overestimated $\hat{\delta}$ because of the inclusion of many intrafamily and forced sales (sheriff's deeds, executor's deeds) whose response to flood zone location differed from those of arm's-length sales.

Taken together, these findings suggest that underreported data processing choices are not inconsequential for the results of hedonic analyses. We acknowledge an important caveat: our illustrative case study is not intended to be representative for all contexts. For instance, hedonic estimates of the location of buildings inside versus outside of discrete flood zone boundaries are likely more sensitive to small spatial errors than estimates of the value of environmental amenities with less discrete spatial variation, such as air pollution or recreational access. Until future empirical work begins to examine the relative importance of those data processing choices in other contexts, a more consistent and transparent reporting in the published literature can help improve shared scrutiny and advance scientific progress.

6. Conclusion

Large-scale property transaction and assessment data offer unprecedented opportunities for detailed empirical research into the dynamics of land ownership, land policy, property valuation, and nonmarket valuation in the United States. After the conclusion of the ZTRAX program, analysts who work with and compile data from county- and state-level data sources, or who purchase similar data services from third-party providers, must address many of the issues discussed herein. Awareness of potential errors and biases, fuller documentation of data processing and filtering choices, and discussion of the potential effects of geographic omissions will enhance the transparency, replicability, and generalizability of empirical findings. We encourage journal editors and referees to require that authors include detailed documentation of data processing choices in their final manuscript submissions. In the absence of official best practices standards, this article can serve as a nonexhaustive checklist of potential issues.

Acknowledgments

Parcel data for approximately two-thirds of U.S. counties were provided by Regrid through its Data with Purpose program (https://regrid.com/purpose). We thank participants of the ZTRAX best data practices workshop (2021 PLACES webinar, June 28–30, 2021) and the ZTRAX workshop for the joint special issues of *Land Economics* and the *Journal of Housing Economics* (August 29–31, 2022) as well as two anonymous reviewers for feedback on prior versions of this article. Christoph Nolte, Adam Pollack, Ido Kushner, and Shelby Sundquist acknowledge support from the Department of Earth and Environment at Boston University, the Junior Faculty Fellows program of Boston University's Hariri Institute for Computing and Computational Science, the Nature Conservancy, and the National Science Foundation's Human-Environment and Geospatial Sciences program (grant 2149243). Johannes Uhl is funded in part by NSF Humans, Disasters and the Built Environment program (grant 1924670). Data are provided by ZTRAX. More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the authors and do not reflect the position of Zillow Group.

References

Albouy, D., P. Christensen, and I. Sarmiento-Barbieri. 2020. "Unlocking Amenities: Estimating Public Good Complementarity." *Journal of Public Economics* 182: 104110. https://doi.org/10.1016/j.jpubeco.2019.104110.

Baldauf, M., L. Garlappi, and C. Yannelis. 2020. "Does Climate Change Affect Real Estate Prices? Only if You Believe in It." *Review of Financial Studies* 33 (3): 1256–95. https://doi.org/10.1093/rfs/hhz073.

Banzhaf, H. S. 2021. "Difference-in-Differences Hedonics." *Journal of Political Economy* 129 (8). https://doi.org/10.1086/714442.

Beltrán, A., D. Maddison, and R. J. R. Elliott. 2018. "Is Flood Risk Capitalised into Property Values?" *Ecological Economics* 146: 668–85.

Bernstein, A., M. T. Gustafson, and R. Lewis. 2019. "Disaster on the Horizon: The Price Effect of Sea Level Rise." *Journal of Financial Economics* 134 (2): 253–72.

Bigelow, D., A. Borchers, and T. Hubbs. 2016. *U.S. Farmland Ownership, Tenure, and Transfer*. Washington, DC: U.S. Department of Agriculture, Economic Research Service.

Bin, O., and J. B. Kruse. 2006. "Real Estate Market Response to Coastal Flood Hazards." *Natural Hazards Review* 7 (4): 137–44.

Bin, O., J. B. Kruse, and C. E. Landry. 2008. "Flood Hazards, Insurance Rates, and Amenities: Evidence from the Coastal Housing Market." *Journal of Risk and Insurance* 75 (1): 63–82.

Bishop, K. C., N. V. Kuminoff, H. S. Banzhaf, K. J. Boyle, K. von Gravenitz, J. C. Pope, V. K. Smith, and C. D. Timmins. 2020. "Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for

Environmental Quality." *Review of Environmental Economics and Policy* 14 (2): 260–81.

Brown, J. F., H. J. Tollerud, C. P. Barber, Q. Zhou, J. L. Dwyer, J. E. Vogelmann, T. R. Loveland, *et al*. 2020. "Lessons Learned Implementing an Operational Continuous United States National Land Change Monitoring Capability: The Land Change Monitoring, Assessment, and Projection (LCMAP) approach." *Remote Sensing of Environment* 238: 111356.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Chaudhry, A. M., D. H. K. Fairbanks, and C. Nolte. 2024. "Water Market Participation and Agricultural Land Values." *Land Economics* 100 (1): 127–47. https://doi.org/10.3368/le.100.1.102122-0091R.

Christensen, G., and E. Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Clapp, C. M., J. Freeland, K. Ihlanfeldt, and K. Willardsen. 2018. "The Fiscal Impacts of Alternative Land Uses." *Public Finance Review* 46 (5): 850–78.

Clarke, W., and M. Freedman. 2019. "The Rise and Effects of Homeowners Associations." *Journal of Urban Economics* 112 (April): 1–15. https://doi.org/10.1016/j.jue.2019.05.001.

Ferraro, P. J., and P. Shukla. 2020. "Feature: Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" *Review of Environmental Economics and Policy* 14 (2): 339–51.

———. 2022. "Credibility Crisis in Agricultural Economics." *Applied Economic Perspectives and Policy* (February): 1–17.

Fraenkel, R. 2020. "Property Tax-Induced Mobility and Redistribution: Evidence from Mass Reappraisals." *Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* 113:1–49.

Friedl, J. 2006. *Mastering Regular Expressions*. Sebastopol, CA: O'Reilly Media.

Gindelsky, M., J. Moulton, and S. Wentland. 2019. "Valuing Housing Services in the Era of Big Data: A User Cost Approach Leveraging Zillow Microdata." *International Conference on Real Estate Statistics*.

Gourevitch, J. D., C. Kousky, Y. Liao, C. Nolte, A. B. Pollack, J. R. Porter, and J. A. Weill. 2023. "Unpriced Climate Risk and the Potential Consequences of Overvaluation in US Housing Markets." *Nature Climate Change* 13: 250–57.

Guignet, D., M. T. Heberling, M. Papenfus, and O. Griot. 2022. "Property Values, Water Quality, and Benefit Transfer: A Nationwide Meta-analysis." *Land Economics* 98 (2): 191–218. https://doi.org/10.3368/le.98.2.050120-0062R1.

Guignet, D., R. R. Jenkins, C. Nolte, and J. Belke. 2023. "The External Costs of Industrial Chemical Accidents: A Nationwide Property Value Study." *Journal of Housing Economics* 62:101954. https://doi.org/10.1016/j.jhe.2023.101954.

Guignet, D., and C. Nolte. 2023. "Hazardous Waste and Home Values: An Analysis of Treatment and Disposal Sites in the U.S." *Journal of the Association of Environmental and Resource Economists* (preprint). https://doi.org/10.1086/726157.

Hanley, N., and M. Czajkowski. 2019. "The Role of Stated Preference Valuation Methods in Understanding Choices and Informing Policy." *Review of Environmental Economics and Policy* 13 (2): 248–66.

Hennighausen, H., Y. Liao, C. Nolte, and A. Pollack. 2023. "Flood Insurance Reforms, Housing Market Dynamics, and Adaptation to Climate Risks." *Journal of Housing Economics* 62:101953. https://doi.org/10.1016/j.jhe.2023.101953.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, *et al*. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59: 944–60. https://doi.org/10.1111/ecin.12992.

Johnston, R. J., and K. Moeltner. 2019. "Special Flood Hazard Effects on Coastal and Interior Home Values: One Size Does Not Fit All." *Environmental and Resource Economics* 74 (1): 181–210. https://doi.org/10.1007/s10640-018-00314-7.

Leyk, S., J. H. Uhl, D. S. Connor, A. E. Braswell, N. Mietkiewicz, J. K. Balch, and M. Gutmann. 2020. "Two Centuries of Settlement and Urban Development in the United States." *Science Advances* 6 (23): eaba2937. https://doi.org/10.1126/sciadv.aba293.

Mamun, S., A. Castillo-Castillo, K. Swedberg, J. Zhang, K. Boyle, D. Cardoso, C. L. Kling, *et al*. 2023. "Valuing Water Quality in the United States Using a National Dataset on Property Values." *Proceedings of the National Academy* *of Sciences* 120 (15): e2210417120. https://doi.org/10.1073/pnas.2210417120.

Mamun, S., E. Nelson, and C. Nolte. 2024. "Estimating the Impact of Critical-Habitat Designation on the Values of Developed and Undeveloped Parcels." *Land Economics* 100 (1): 148–75. https://doi.org/10.3368/le.100.1.101922-0081R.

Maniadis, Z., F. Tufano, and J. A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104 (1): 277–90.

———. 2017. "To Replicate or Not to Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study." *Economic Journal* 127 (605): F209–35. https://doi.org/10.1111/ecoj.12527.

Manson, S., J. Schroeder, D. Van Riper, and S. Ruggles. 2018. "IPUMS National Historical Geographical Information System." Version 13.0. Available at https://www.nhgis.org (accessed May 14, 2020).

Microsoft. 2018. "U.S. Building Footprints." Available at https://github.com/microsoft/USBuildingFootprints (accessed February 2, 2020).

Moulton, J. G., N. J. Sanders, and S. A. Wentland. 2018. "Toxic Assets: How the Housing Market Responds to Environmental Information Shocks." *Land Economics* 100 (1): 66–88. https://doi.org/10.3368/le.100.1.102122-0089R.

Murfin, J., and M. Spiegel. 2020. "Is the Risk of Sea Level Rise Capitalized in Residential Real Estate?" *Review of Financial Studies* 33 (3): 1217–55. https://doi.org/10.1093/rfs/hhz134.

Netusil, N. R., K. Moeltner, and M. Jarrad. 2019. "Floodplain Designation and Property Sale Prices in an Urban Watershed." *Land Use Policy* 88 (November): 104112. https://doi.org/10.1016/j.landusepol.2019.104112.

Nolte, C. 2020. "High-Resolution Land Value Maps Reveal Underestimation of Conservation Costs in the United States." *Proceedings of the National Academy of Sciences* 117 (47): 29577–83. https://doi.org/10.1073/pnas.20128651.

Olofsson, P., C. E. Holden, E. L. Bullock, and C. E. Woodcock. 2016. "Time Series Analysis of Satellite Data Reveals Continuous Deforestation of New England since the 1980s." *Environmental Research Letters* 11 (6). https://doi.org/10.1088/1748-9326/11/6/064002.

Pollack, A. B., D. H. Wrenn, C. Nolte, and I. Sue Wing. 2023. "Potential Benefits in Remapping the Special Flood Hazard Area: Evidence from the U.S. Housing Market." *Journal of Housing Economics* 61: 101956. https://doi.org/10.1016/j.jhe.2023.101956.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Swedberg, K., D. S. Cardoso, A. Castillo-Castillo, S. Mamun, K. J. Boyle, C. Nolte, M. Papenfus, and S. Polasky. 2024. "Spatial Heterogeneity in Hedonic Price Effects for Lake Water Quality." *Land Economics* 100 (1): 89–108. https://doi.org/10.3368/le.100.1.102122-0086R.

Uhl, J. H., S. Leyk, C. M. McShane, A. E. Braswell, D. S. Connor, and D. Balk. 2021. Fine-Grained, Spatiotemporal Satasets Measuring 200 Years of Land Development in the United States. *Earth System Science Data* 13: 119–53. https://doi.org/10.5194/essd-13-119-2021.

U.S. Census Bureau. 2021. "U.S. Census Geocoder." Available at https://geocoding.geo.census.gov/geocoder (accessed June 15, 2021).

U.S. Geological Survey. 2017. "National Hydrography Dataset." Available at https://nhd.usgs.gov/NHD_High_Resolution.html (accessed August 29, 2019).

Wentland, S. A., Z. H. Ancona, K. J. Bagstad, J. Boyd, J. L. Hass, M. Gindelsky, and J. G. Moulton. 2020. "Accounting for Land in the United States: Integrating Physical Land Cover, Land Use, and Monetary Valuation." *Ecosystem Services* 46: 101178. https://doi.org/10.1016/j.ecoser.2020.101178.

Zabel, J., C. Nolte, and R. Paterson. 2024. "Measuring the Value of U.S. National Parks Using Hedonic Property Value Models." *Land Economics* 100 (1): 176–99. https://doi.org/10.3368/le.100.1.102122-0088R1.

Zhang, J., D. J. Phaneuf, and B. A. Schaeffer. 2022. "Property Values and Cyanobacterial Algal Blooms: Evidence from Satellite Monitoring of Inland Lakes." *Ecological Economics* 199: 107481. https://doi.org/10.1016/j.ecolecon.2022.107481.

Zillow Group. 2021a. "ZTRAX: Frequently Asked Questions." Available at https://www.zillow.com/research/ztrax/ztrax-faqs.

———. 2021b. "Zillow's Transaction and Assessment Database (ZTRAX)." Available at http://www.zillow.com/research/ztrax.