Studying the impacts of environmental amenities and hazards with nationwide property data: best data practices for interpretable and reproducible analyses

Christoph Nolte[1,2,*], Kevin J. Boyle[3,4], Anita Chaudhry[5], Christopher Clapp[6], Dennis Guignet[7], Hannah Hennighausen[8], Ido Kushner[1], Yanjun Liao[9,10], Saleh Mamun[11,12], Adam Pollack[1], Jesse Richardson[13], Shelby Sundquist[1], Kristen Swedberg[3], Johannes H. Uhl[14,15]

[1] *Department of Earth & Environment, Boston University, Boston, Massachusetts, United States of America*
[2] *Faculty of Computing & Data Sciences, Boston University, Boston, Massachusetts, United States of America*
[3] *Department of Agricultural and Applied Economics, Virginia Tech, Blacksburg, Virginia, United States of America*
[4] *Program in Real Estate, Virginia Tech, Blacksburg, Virginia, United States of America*
[5] *Department of Economics, California State University, Chico, California, United States of America*
[6] *Harris School of Public Policy, University of Chicago, Chicago, Illinois, United States of America*
[7] *Department of Economics, Appalachian State University, Boone, North Carolina, United States of America*
[8] *Department of Economics, University of Alaska, Anchorage, Arkansas, United States of America*
[9] *Wharton Risk Center, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America*
[10] *Resources for the Future, Washington, District of Columbia, United States of America*
[11] *Department of Applied Economics, University of Minnesota, St. Paul, Minnesota, United States of America*
[12] *Natural Resources Research Institute, University of Minnesota, Duluth, Minnesota, United States of America*
[13] *College of Law, West Virginia University, Morgantown, West Virginia, United States of America*
[14] *Institute of Behavioral Science, University of Colorado, Boulder, Colorado, United States of America*
[15] *Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, United States of America*

** Corresponding author. Email: chrnolte@bu.edu (CN)*

Abstract

Access to rich, nationwide property data has catalyzed rapid empirical work concerning land use choices in several fields of inquiry, including environmental economics, urban geography, and conservation biology. When data on property transactions and assessments are provided in its original or only partially pre-processed state, the accuracy, reliability, and generalizability of findings can be improved with a series of cleaning procedures and quality checks. We discuss issues inherent in using increasingly popular, nationwide data to perform econometric analyses and propose best practices for data preparation by example of ZTRAX, a U.S.-wide real estate database available to academics, non-profit, and government researchers between 2016 and 2023. We cover (1) the identification of arms-length sales, (2) the geo-location of parcels and buildings, (3) temporal linkages between transaction, assessor, and parcel data, (4) the identification of property types, such as single-family homes and vacant lands, and (5) dealing with missing or mismeasured data for standard housing attributes. We provide supplementary maps, filtering tables, and algorithmic descriptions to help analysts check and document their choices, improve the quality of ongoing and planned research, and help readers better understand the scope, reliability, and generalizability of findings and data products.

1. Introduction

We discuss issues inherent in using increasingly popular, nationwide property data to perform statistical analyses, and propose best practices for data preparation in the United States. By

example of Zillow's ZTRAX, a U.S.-wide property database, we draw attention to common issues that, if uncorrected or undocumented, can bias research findings and hinder reproducibility. These issues are of particular importance in nationwide analyses where data generation is decentralized, interpretation nuanced, and for which no published codebooks and "best practice" guidelines exist. We propose concrete, interpretable, and reproducible solutions, and share machine-readable tables and code to facilitate analyses of ZTRAX data. In making this information available, this document also serves as a technical reference for our future analyses using large-scale property and transaction data in the United States.

ZTRAX is a U.S.-wide real estate database managed by Zillow Inc., an American online real estate marketplace company (Zillow 2019). In 2016, Zillow made two-year research licenses to ZTRAX data available free of charge to U.S. academic, non-profit, and government researchers. As of June 2021, ZTRAX contained tax assessor data on property characteristics, geographic information, and valuation for approximately 150 million parcels in over 3,100 U.S. counties, alongside more than 400 million public transaction records in more than 2,750 U.S. counties. The release of such a large real-estate database free of charge radically lowered the barriers of entry for empirical analyses and catalyzed rapid growth in published work, often at unprecedented scales and spatial resolutions. Peer-reviewed ZTRAX-based publications have contributed to domains as diverse as the geography of urban growth (Leyk & Uhl 2018; Connor et al. 2020; Leyk et al. 2020; Uhl et al. 2021a), urban economics (Clarke & Freedman 2019; Peng & Zhang 2019), national land accounting (Wentland et al. 2020), impacts of land use regulations (Onda et al. 2020; Leonard et al. 2021), residential choices on energy (Shen et al. 2021) and water use (Quesnel et al. 2020), responses to fluvial flooding (Pinter & Rees 2021) and sea level rise (Bernstein et al. 2019; Buchanan et al. 2020; Murfin & Spiegel 2020), the valuation of park safety (Albouy et al. 2020), lake water quality (Moore et al. 2020), and neighborhood demographics (Zhang & Leonard 2021), as well as the estimation of land values and conservation costs (Nolte 2020).

In June 2021, Zillow announced that it would stop accepting applications for new two-year ZTRAX licenses on July 15th, 2021, and that access to ZTRAX data would end on Sept 30, 2023. This leaves researchers with a unique opportunity to develop insights and data products from ZTRAX within the remaining two-year period. The short time frame also adds pressure to publish results quickly and could reduce time spent on data scrutiny and validation. Awareness of potential errors and biases, the documentation of filtering choices, and the discussion of the potential effects of geographic omissions can reduce the possible influence of "hidden" researcher decisions (Huntington-Klein et al. 2021) and can enhance the quality, validity, generalizability, and interpretability of work produced over the next two years. As we expect these issues not to be unique to ZTRAX, this document can also serve a guide to data quality considerations for researchers working with similar datasets from other data providers, and for analysts aggregating property and transaction data across multiple sources and locations.

The choice of topics is shaped by our shared interests in the empirical estimation of the magnitude and distribution of impacts of environmental amenities and hazards on the value of properties. Contributors to this document are currently using ZTRAX data across multiple independent analyses, in which estimates of interest include: the cost of land acquisitions for conservation purposes; the benefits of lake water quality and the cost of its impairment; the

risk of flood damage to residential homes; the effects of flood insurance policies; the cost of hazardous waste incidents and the benefits of subsequent cleanups; and property value impacts of critical habitat under the U.S. Endangered Species Act. Through this work, we have identified common problems of working with ZTRAX data, and experimented with potential solutions in the following areas:

1. Identifying transaction prices reflecting fair market value
2. Geo-locating transacted properties: land and buildings
3. Linking transactions to time-variant property characteristics
4. Identifying different types of properties, e.g., single-family homes and vacant lands
5. Dealing with missing or mismeasured data for standard housing attributes

In the following sections, we introduce each issue, establish its relevance, and propose solutions. We see this document as a starting point for a development of best data practice guidelines for property-based analyses, using ZTRAX as an illustrative example. 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, we hope this document will help (i) promote high-quality research, (ii) understand limitations to findings, including by referees who do not have access to ZTRAX, and (iii) facilitate replication.

2. Background and data

The contents of this document are the product of a group effort by academic researchers from ten U.S. universities who use ZTRAX data in a variety of property-level analyses, and have a shared interest in accuracy, reliability, and reproducibility. Many of the insights shared below have been obtained by validating ZTRAX records against supplementary datasets, such as digital parcel maps and building footprints. The authors affiliated with Boston University have linked most ZTRAX records for the contiguous United States (CONUS) to county-level digital parcel maps using text-based tax assessor parcel identifiers and conversion algorithms developed as part of the Private-Land Conservation Evidence System (PLACES) (Nolte 2020) and described below. Because two thirds of U.S.-wide digital parcel maps in PLACES are licensed from third-party providers, we cannot publicly share the full parcel-level dataset underlying our claims, such as the corrected parcel and building coordinates. However, with the methodological descriptions we offer below and access to similar supplementary data, the computationally versed reader will be able to reproduce our findings for their study region and implement the proposed corrections and filters. Unless otherwise noted, findings in this paper are based on ZTRAX data downloaded from Zillow's servers on Feb 3, 2020 and limited to CONUS.

2.1. An overview of ZTRAX

According to the ZTRAX FAQ, Zillow sources ZTRAX "from a major large third-party provider and through an internal initiative we call County Direct" (Zillow 2021).

ZTRAX consists of three databases. First, a property transaction database ("ZTrans") contains >400 million public transaction records including deeds, mortgages, and foreclosures in a

relational database structure with 21 tables and 470 fields. Findings and filters presented in this paper are based on data from four tables: "Main" (prices, dates, document type, partial interest codes, intra-family flags), "PropertyInfo" (assessor parcel numbers), "BuyerName", and "SellerName" (names of buyers and sellers).

Second, a tax assessment database ("ZAsmt") stores property-level records extracted from property tax roll data in a relational database structure with 23 tables and 352 fields. Because property tax assessors are tasked with maintaining a complete account of the taxable value of all properties within their jurisdiction, ZAsmt can usually be expected to contain the full list of properties within a given county or town. However, the set of variables collected for each property varies substantially across geographies (see below). Most ZTRAX-based analyses will use only a small number of fields. Findings and filters presented in this manuscript are based on data from four tables: "Main" (assessor parcel numbers, tax account identifiers, lot size, geo-coordinates, tax year), "Building" (building characteristics, land use codes), "Value" (tax valuation, estimated fair market value), and "Name" (owner names). ZTrans and ZAsmt can be linked through a unique property identifier provided by ZTRAX ("ImportParcelID"), which appears to have been derived from (text-based) assessor parcel numbers and links the two databases with very high, though not perfect, accuracy.

Third, ZTRAX contains historical versions of the ZAsmt database ("historical ZAsmt") that allows the tracking of changes to the database at the property record level. The temporal coverage of historical ZAsmt extends back to the early 2000s in most states but varies across geographies (Fig S1).

Geographic coverage of ZTrans transaction data (>2,750 counties) is more limited than that of ZAsmt property characteristics (>3,100 counties). Geographic differences are particularly pronounced in the availability of price information, the dependent variable of interest in our revealed-preference studies. The availability of price information is strongly shaped by the extent to which U.S. states require disclosure of sales prices (Fig 1). Lists of non-disclosure states commonly include: Alaska, Idaho, Indiana, Kansas, Louisiana, Mississippi, Montana, New Mexico, North Dakota, South Dakota, Texas, Utah, and Wyoming (e.g., Wentland et al. 2020). We also find sales price data to be relatively rare in Maine and Missouri, as well as 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 scrutiny.

States and counties also vary in the length of the time period for which transaction data is consistently available. Most counties contain transaction data for the first two decades of the 21st century. Three or more decades of data are available in Northeastern states (Connecticut, Maryland Massachusetts, New Jersey, New York, Rhode Island), much of California, Florida, Tennessee, and Ohio, as well as urban centers across the country (Fig S2).

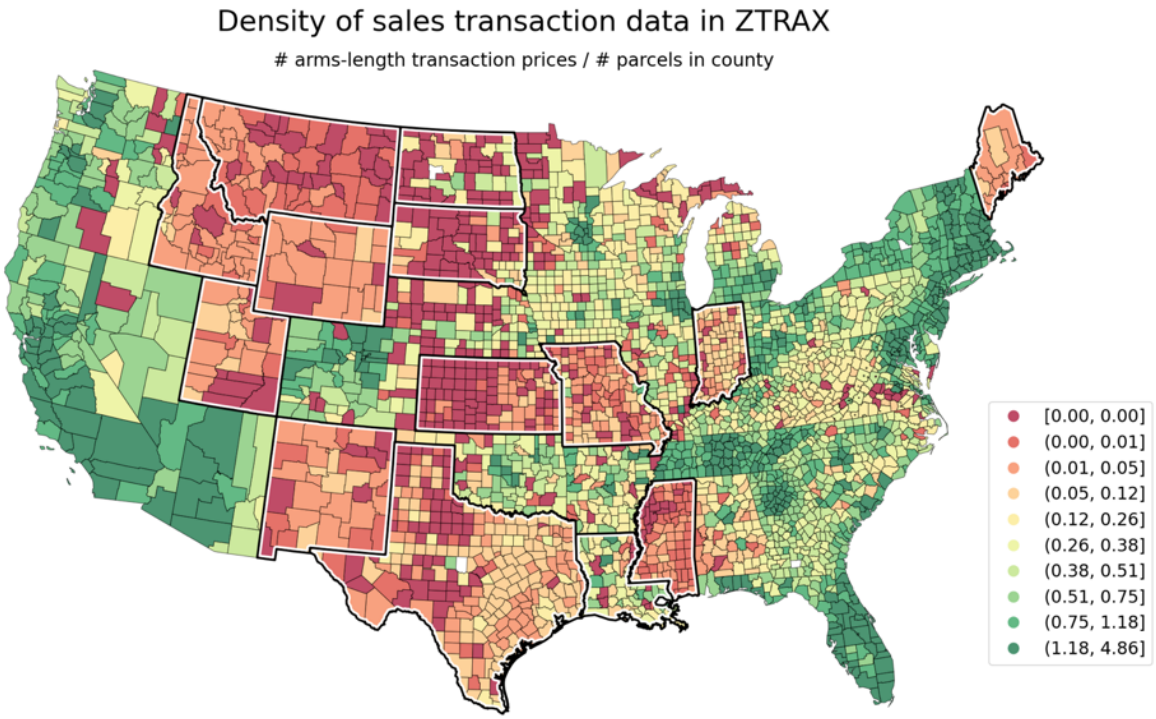


Fig 1: Density of sales transaction data in ZTRAX. Density is computed as the county-level ratio of (i) the count of non-duplicate arms-length transaction records with prices >$1000 and (ii) the count of parcels in digital parcel maps. Black/white boundaries show non-disclosure states or states with unusually scarce price data. Note that the time period spanned by ZTrans transaction data varies across counties (Fig S2), which contributes to the observed differences in sales densities.

3. Challenges

3.1. Identifying transaction prices reflecting fair market value

Relevance: Real estate appraisals, hedonic pricing methods, risk assessments, and other property analyses often rely on the assumption that sales prices of property transactions 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, neither being under any compulsion to buy or sell. Transactions for properties at prices other than FMV are often observed in public record datasets, including ZTrans. Non-FMV price observations include transactions between family members; transactions under distress (such as foreclosures); transactions below market value from public actors (e.g., by a targeted sale to veterans); or listed prices referring to monetary amounts other than the full property value (e.g., loans, mortgages, partial interests). To avoid bias that can affect subsequent conclusions, researchers need to be able to identify and isolate fair market value transactions.

Problem: Guidelines on how to filter ZTRAX for transactions that reflect fair market value are limited. ZTrans contains several fields that can aid with the filtering for fair market value transactions. Zillow advises that "extensive exploring on [the part of analysts] is required" (Zillow 2021). The interpretation of fields often requires domain expertise, such as on the legal meaning and usage of document types, which can vary by state. Ancillary variables that seem to address the above concerns are often incomplete. For instance, ZTrans contains an "intra-family

transfer flag", which identifies transactions between family members, but the algorithm is not documented. Based on a comparison of buyer and seller names (see below), we estimate that this flag misses potentially 15.2 million intra-family transactions (6.1% of deed records). Zillow's FAQ confirm that their own cleaning procedure includes an internal text-matching algorithm (Zillow 2021) that is not publicly documented.

Potential Solutions: We propose filters for 9 ZTRAX variables to identify transactions whose prices are more likely to reflect FMV (Supporting Information). 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 resulting confidence levels can be flexibly combined. As an example, an analyst could choose to aggregate filters to the lowest confidence level for each given transaction, and then only retain "high confidence" transactions that obtained a "high confidence" value across all filters.

Six of the nine listed filters are straightforward exclusions based on discrete values (data types, document types, loan types, price types, intra-family flags) or cutoffs (for token prices). Three warrant further elaboration:

1. To enhance the identification of intra-family transfers, we compute a **similarity index between buyer and seller names**. Running at the county-level, the 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 county. This inverse frequency weighing reduces the probability that name similarity is erroneously established from frequently occurring words (e.g., "John" or "Michael"). We omit words with ≤2 letters and remove frequently used generic words (e.g., "Bank", "LCC"). Transactions with a similarity index of ≥50% (applied to the maximum of the buyer's and seller's names) are flagged as "likely" intra-family transfers. Using this method, we estimate that a naïve approach to identifying intra-family transactions based on document codes and the intra-family transfer flag underestimates the actual extent of intra-family transfers. Of 245 million deed records, ZTRAX flags 23.0% with its intra-family transfer flags. We estimate the actual number to be 29.1%, a difference of 15.2 million transactions.
2. Developing filters for the 161 standardized **document type** categories is not trivial. Document types can have different meanings and usages in different states. For instance, warranty deeds (WRDE) are the most frequent source of FMV transactions in most states but grant deeds (GRDE) are more frequent in California, Nevada, and Vermont. Quitclaim deeds (QCDE) rarely contain sales price information in most states except in Massachusetts and Vermont, where prices are frequently provided (Fig S3). Because ZTrans contains 3,837 unique combinations of states and document types, an in-depth assessment of each combination is infeasible. Our filters therefore combine a hierarchical exclusion filter with a data-driven follow-on: First, with the help of a land use attorney, we make deterministic choices for the most frequent unambiguous document codes, including flags for foreclosures, intra-family transfers, loans, and cancellations. An explanation of the meanings of the most frequent deeds is provided in the Supporting Information. For the remaining, non-excluded 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 >$1000, and 2) the percentage of non-family transactions (see above). We flag state-code combinations where at least one of these statistics was found to be <10% or <33% as "low" and "medium" confidence, respectively.

3. We identify **public buyers and sellers** from their respective name fields using string pattern matching (regular expressions). There is a broad range of public organizations in the United States, and their names can be spelled in a diverse range of ways within and across county registries. In states and regions where price data is scarce (e.g., Arkansas, Indiana, Texas), a large share of sale observations comes from public sources of public sale data, which can bias estimates of property value downwards. 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 therefore best described as a hybrid of top-down (names of pre-identified agencies) and bottom-up (spelling learned from ZTrans) identification of string patterns of public organizations. Due to the absence of an external validation dataset for testing, we do not claim that the set of expressions is comprehensive. Instead, we share our table of regular expressions (Supporting Information) and will post feedback we receive from the community of ZTRAX researchers on http://placeslab.org/ztrax.

3.2. Geo-locating transacted properties: land and buildings

Relevance: Property and real estate analyses, especially those studying relatively local spatial phenomena, rely on accurate information of the location of land and buildings. For example, hedonic property value analyses often infer landowners' preferences for environmental characteristics from spatial associations between the prices of transacted parcels and environmental variables of interest. To do so, analysts need to geo-locate each transaction, and then computationally derive variables of interest from spatial data of environmental attributes (vector or rasters). Depending on the application, demands on spatial precision can be high. For instance, Netusil et al. (2019) show that estimates of impacts of floodplain location on property values can be very sensitive to the choice of parcel boundaries vs. buildings footprints as the spatial reference. Many analyses also benefit from information on the characteristics of the land under each parcel. For example, the estimation of the impacts of development restrictions (or the cost of conservation easements) benefits from area-based proxies of a parcel's potential for future development, which in turn is affected by terrain, wetlands, flood risk, existing land cover, etc. Point locations are usually insufficient for the computation of area-based proxies. Instead, information of the actual location of the parcel boundary (polygon) is often required.

Problem: ZTRAX does not contain parcel boundary data. Many ZTRAX-based spatial analyses rely on point locations (latitude and longitude) that can be found in tax assessor (ZAsmt) and transaction records (ZTrans). Zillow's FAQ describe these data as "enhanced Tiger coordinates" (Zillow 2021) and "Populated by GEOCoder" (ZTRAX documentation), which might refer to the U.S. Census Geocoder (U.S. Census Bureau 2021). Using these coordinates without careful attention to coordinate system projection, duplicates, missing data, and building locations can lead to misleading or unrepresentative findings. Six issues are most worthy of attention (Fig 2):

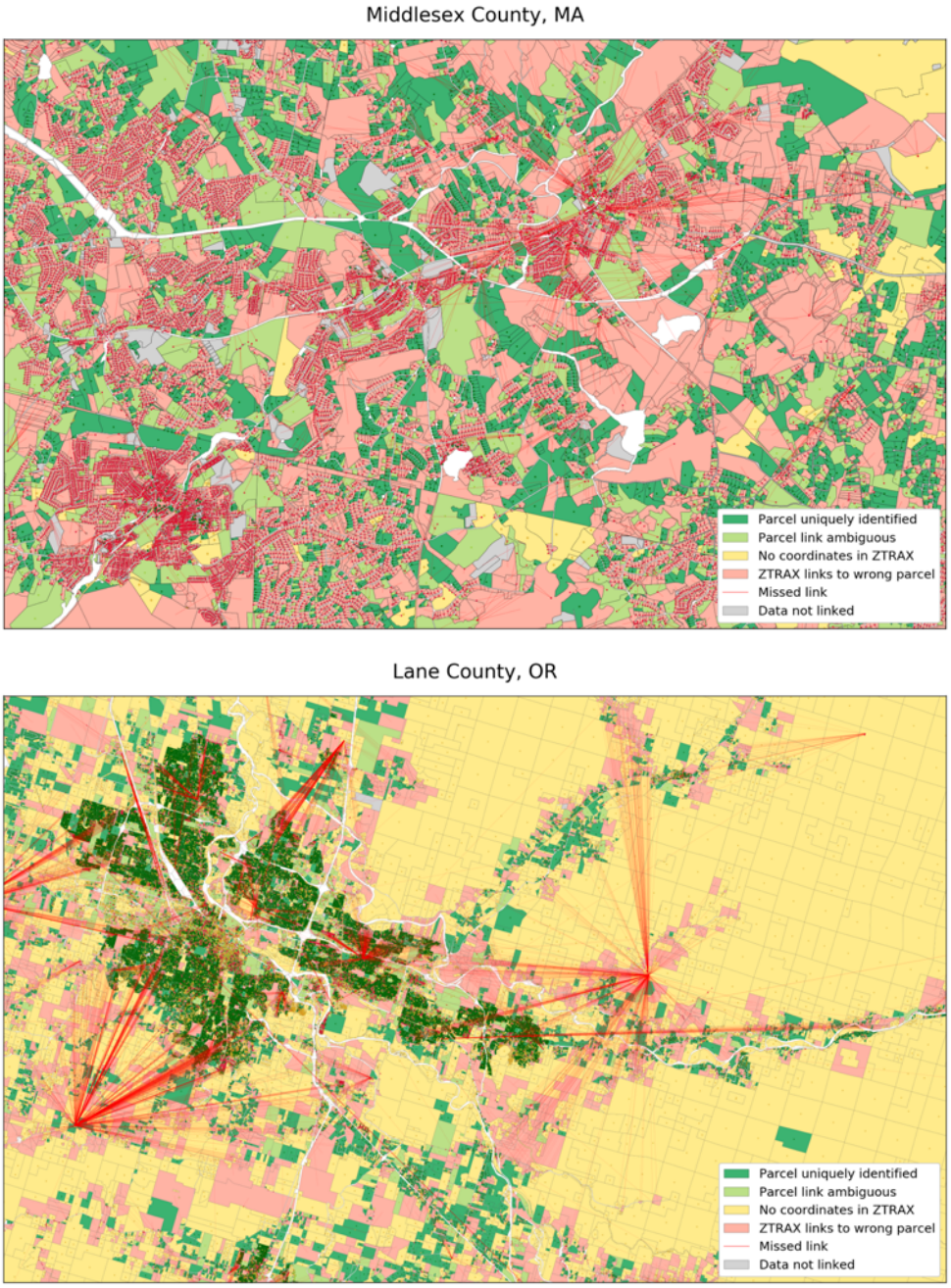


Fig 2: Illustration of potential geo-location errors of ZTRAX data by example of the town of Concord in Middlesex County, Massachusetts (top) and the city of Eugene in Lane County, Oregon (bottom). Errors shown would occur if the analyst made the (naïve) assumptions that ZTRAX coordinates are correct, based on the WGS84 datum and that they can be used to identify parcels polygons in digital maps. In Middlesex County, which uses the NAD27 datum, this approach would link up the vast majority of urban parcels incorrectly (dark red points) due to a translation error of approximately 40 meters. Most large, rural parcels would have been either incorrectly (red) or ambiguously linked (light green: multiple coordinates on one parcel, only one of which is correct). In Lane County, which predominantly uses the WGS84 datum, the analyst would correctly and uniquely identify a majority of urban parcels (dark green points). However, most non-urban parcels have no parcel coordinates (yellow). Furthermore, a non-trivial number of urban and rural parcels appear to be incorrect and derived from ZIP code area centroids (red lines show distance between correct centroid and ZTRAX coordinates and converge around a small number of points).

1. ZTRAX latitude and longitude coordinates appear to have been derived using various **geodesic datums** (e.g., NAD27, NAD83, WGS84), but the datum information is not contained in the database. Comparisons with geo-referenced parcel boundary data suggest that the predominant datum varies by county in a non-predictable pattern (Fig S4). We also find counties using multiple datums. Not correcting for these errors can lead to systematic geo-location errors that vary in magnitude across geographies. Errors will generally be higher on the coasts (~100m in California, ~40m in New England states), but are largely ignorable in the Midwest (e.g., Indiana and Michigan).
2. Some ZTRAX coordinates seem to have been derived from **ZIP code area centroids** instead of parcel data. These cases are not flagged in the database. Using these coordinates can lead to geo-location errors well above 1km.
3. Many records are **lacking point locations** (Fig S5). Missing coordinate data is often associated with particular types of parcels (e.g., vacant parcels, rural parcels, records without parcel addresses). Excluding them from an analysis will usually have non-random effects on sample characteristics.
4. In most counties, ZTRAX coordinates are based on **parcel** centroids. In others, they can refer to **building** locations. ZTRAX provides no information to distinguish between these two cases. Distances between building footprints and parcel centroids vary across the country (Fig S6); mean distances of >100m are common in rural settings with large parcels. Analyses that require precise geo-location of buildings are subject to errors of possibly large magnitudes. Major effects on findings have been reported when assessing the impact of spatially precise policies (legal floodplains) (Netusil et al. 2019). Uhl et al. (2021) compared ZTRAX locations to remote-sensing derived building footprint data (Microsoft 2018) and find positional accuracy to decrease from urban to rural settings.
5. Most counties contain at least some **incorrect, non-duplicate** parcel locations (Fig S7). Possible observed reasons include coordinates that are based on owner's mailing addresses (instead of property location addresses), as well as subdivisions of parcels.
6. Point locations can **change between versions** of ZTRAX. 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 to be replaced with street address geo-locations (Fig S8). For many Rhode Island properties, we also found minor shifts in point locations, which were multidirectional and thus not simply attributable to changes in projection (Fig S8). While such changes appear to reflect improvements in geo-location over time, they also highlight the need to exercise particular caution in geo-location records when leveraging tax assessor datasets from multiple time periods.

If ZTRAX coordinates are derived from original datasets that find widespread use (e.g., the U.S. Census Geocoder), these issues will likely not only occur in ZTRAX data, but also in other uncorrected property datasets derived from tax assessor and address location data.

Potential Solutions: There are several options to improve the geo-location of ZTRAX data. Because of observed improvements in geo-coordinates over time, we recommend starting with the most recent ZTRAX version available. Analyst can then choose among the following options

as a function of their resource constraints (data access and time available for data inspection and cleaning) and the expected sensitivity of their findings to geo-location errors.

- **Quick fixes**. Analysts lacking access to digital parcel maps or the geoprocessing skills to link ZTRAX records to parcels and buildings can enhance the reliability of their findings with two fixes. Specifically, we recommend (i) ensuring that the correct datum is used to spatially locate the ZTRAX coordinates (Fig S4, Supporting Information) and (ii) dropping entries with duplicate ZTRAX coordinates, potentially combined with verifying if dropped coordinates are ZIP code area centroids or if dropped records have no or non-duplicate street addresses. Fig S5 shows the prevalence of missing and non-empty duplicate coordinates in ZAsmt and, therefore, of anticipated reductions in sample size and county coverage when removing these entries. Fig S7 shows how much geo-location error remains in each county after implementing these two fixes: in most counties, the median geo-location error drops to below <1m, suggesting that most parcels will be correctly located. However, mean errors can remain 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**. County identifiers (all records) and ZIP codes (most records) are independently provided and can be used to remove coordinates that fall outside the corresponding spatial boundaries. Official digital boundaries of counties and ZIP codes are available through IPUMS (Manson et al. 2018). We do not recommend cropping to census tract or block boundaries, as the identifiers for census units appear to have been derived directly from the geo-coordinates through spatial operations.
- **Geocode addresses**. Many records with missing geo-coordinates contain addresses (street, city, and zip code) (Fig S9). Analysts could potentially improve the completeness of geolocations in ZTRAX by means of additional geocoding, e.g., by using the U.S. Census Geocoder or the GeoCoder API (https://geocoder.readthedocs.io). This approach remains untested and potentially vulnerable to the same issues as observed for ZTRAX coordinates. Analysts will likely still have to apply the quick-fix filters listed above.
- **Linking ZTRAX to spatial parcel boundaries**. Geo-referenced digital parcel maps now exist in at least 3,073 (97.8%) of U.S. counties. In most cases, they can be uniquely and reliably linked to ZTRAX tax assessor data using unique parcel identifiers (assessor parcel numbers, APN) or, in some cases, unique taxpayer account numbers. This approach will generally lead to a more reliable and complete geo-location of ZTRAX than quick fixes. It also allows the analyst to derive precise spatial indicators directly from the polygon data (e.g., road access, water frontage, area-based measures). However, establishing this linkage is complicated by idiosyncratic differences in the syntaxes of APNs, which vary across ZTRAX' tax assessor and digital parcel datasets, as well as by county or, in the Northeast, by town. We recommend using pattern-based text extraction (regular expressions) and post-processing of extracted strings in order to empirically identify the translation function that leads to the highest number of uniquely identified ZTRAX-parcel matches in each county or New England town. Using this approach, analysts will be able to link most parcel boundaries to ZAsmt records in most counties (Fig S10).

- **Identifying building locations**. After linking ZTRAX records to parcel polygons, analysts can use spatial data on building footprints to identify the location of buildings within a parcel. A U.S.-wide open-source dataset of 130 million building footprint polygons is available from Microsoft (2018) with updates in 2020. Derived from high-resolution satellite imagery with a documented machine learning algorithm, the dataset is, to our knowledge, the most consistent U.S.-wide indicator of building presence available free of charge. Dates of observation are not provided and can be more than a decade old in some instances. We also observe an underreporting of buildings under tree cover. However, in our analyses, we found building footprints to be the most consistent nationally available open-source indicator for the presence of a building, and use it as a reference dataset against which to compare other indicators of buildings presence (see Section 3.4: "Identifying different types of properties" as well as Uhl et al. 2021b).

3.3. Linking transactions to time-varying property characteristics

Relevance: Analyses often need to establish a reliable link between the price of transactions and the characteristics of the property at the time of sale. Property tax assessments (ZAsmt) are an important public data source for these attributes, often reporting building square footage, architectural style, counts of units, rooms, bedrooms, bathrooms, as well as the presence of other features (garage, pool, etc.). Linked digital parcel maps can provide further information on lot size, building location, land cover, as well as access to roads, water bodies, or open space. These characteristics can change over time as buildings are built, remodeled, or leveled, and as boundaries are redrawn following subdivisions or mergers. Analysts usually want to be certain that characteristics observed in tax assessor data and digital maps are the same as those corresponding to the time period for an observation (e.g., at the time of sale).

Problem: Tax assessor datasets and digital parcel maps provide cross-sectional snapshots of property conditions at one point in time, typically a recent one. Dataset versions for multiple years sometimes exists, as ZTRAX contains past versions of its tax assessor data, and some digital parcel map providers offer archives of historic data. However, synthesizing datasets from multiple time periods substantially increases data volumes and time cost for a given analysis, often with uncertain benefits. Furthermore, not all regions have historical data.

Analysts thus 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 (Fig 3, see also Leyk et al. 2020). We also observed building year values for identical properties to differ between historical and current ZTRAX versions in bidirectional and idiosyncratic ways that cannot be explained by new constructions alone (Fig S11), but indicate that building year data is collected, updated, and interpreted in different ways in different counties. 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 external detection of changes, but come with their own set of challenges, such as classification error, insufficient resolution, or limited temporal coverage.

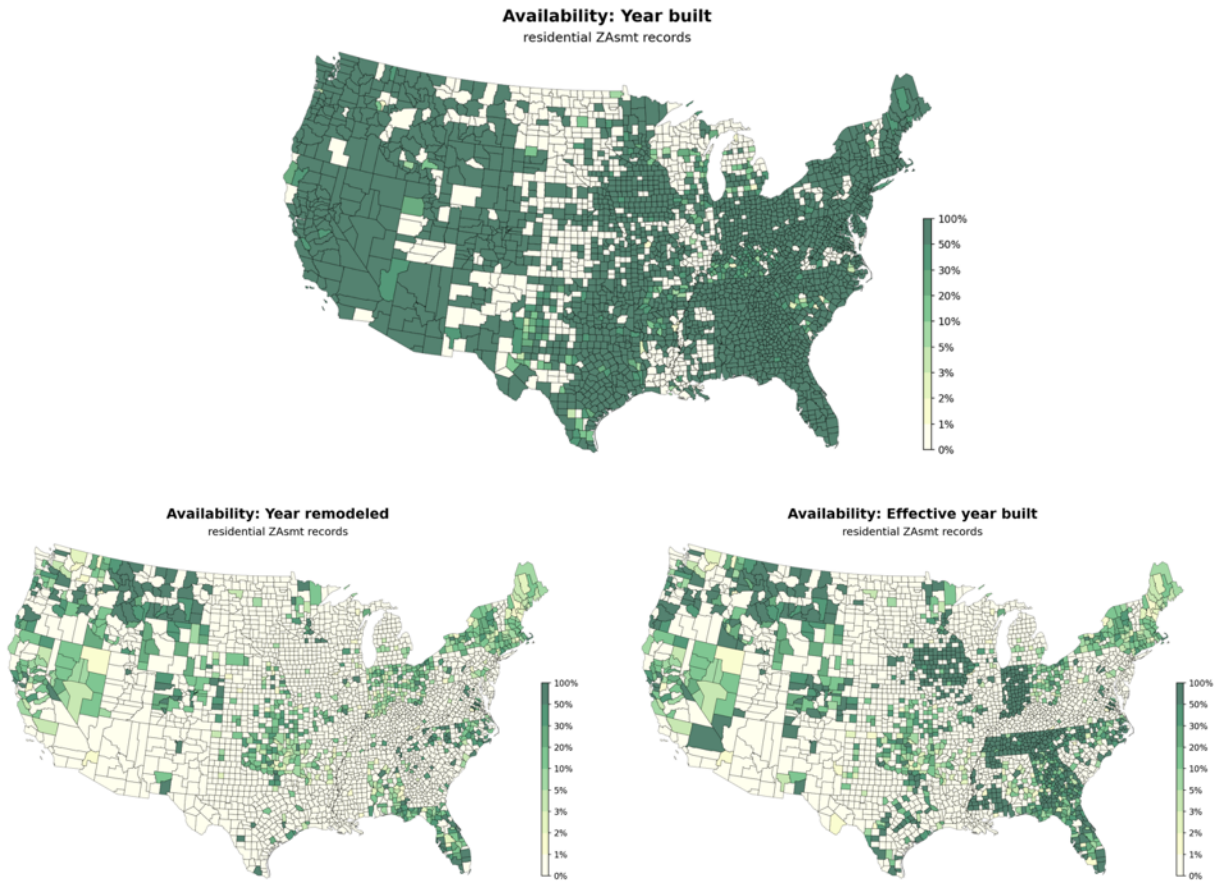


Fig 3: Presence of data for three years of change to buildings in residential (RR) ZAsmt property records. "Year built" (top), "year remodeled" (bottom left) and "effective year built" (bottom right). "Effective year built" could be a hybrid of "built" and "remodeled" year but might also include other adjustments.

Potential Solutions: 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.** Tax assessor data often contains information on (i) the year the building was first constructed and, in a minority of counties, (ii) the year of the last remodeling (Fig 3). 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 (Fig S12). There are two possible approaches to account for these transactions: (a) running the analysis after excluding such observations, and (b) 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 is available, the analyst will also need to 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 (Fig S11), 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, an USGS product that tracks change to land cover from 1985-2017, we find low correspondence between building years and remotely sensed transitions from undeveloped to developed land cover (Fig S13). Modern high-resolution satellites with more frequent temporal coverage (Sentinel-2, Planet Labs) will likely help improve observations of change. However, we do not anticipate nationwide change products to be ready in time for ZTRAX-based analyses. Due to the observed uncertainties associated with this approach, we recommend its use only as a robustness check.
- **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.
- **Using datasets from multiple time periods**. ZTRAX makes previous versions of its assessor database available, and some data aggregators archive historical versions of parcel boundary data. We have not systematically assessed the availability and quality of historical data across the country, but we anticipate that both vary geographically 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).

3.4. Identifying different types of properties

Relevance: 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 (Bishop et al. 2020). Similarly, efforts to estimate the value of undeveloped land (e.g., Nolte 2020) need to be able to reliably identify and exclude parcels with buildings, as buildings often represent a large share of a property's value.

Problem: The availability, usage, and quality of ZTRAX 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 the type of properties identified with a given code varies, often across state boundaries (Fig 4). For instance, in most counties, single-family homes are identified as "RR101", but in other counties, "RR999" (inferred single-family), "RR000" (general residential), or "RR102" (rural residence) are 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).

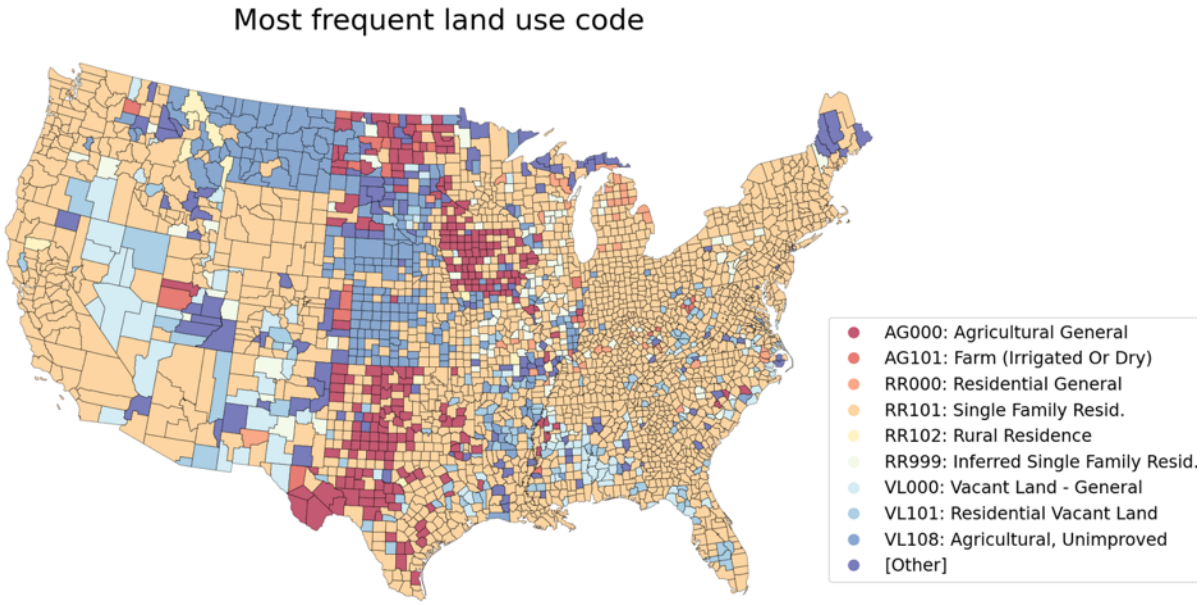


Fig 4: Most frequent land-use code across all parcels linked to ZAsmt in each county.

The identification of vacant lands in ZTRAX is particularly challenging. For instance, the indicator "number of buildings" misses most buildings in hundreds of counties (Fig S14). 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 (Fig S15). Alternative indicators for building presence can be derived from ZTRAX (e.g., property land-use codes, or the assessed value of improvements), or from remote sensing data linked to parcel boundaries (e.g., building footprints, developed land cover). However, no single indicator is unambiguously perfect for a nationwide analysis.

Proposed Solutions: We recommend that analysts of ZTRAX data exercise particular caution in developing their submarket filters, test the robustness of their results to alternative plausible filtering conditions, and document their choice of filtering steps alongside published results.

- **Single-family homes** are likely best identified by combining several land-use codes (RR000, RR101, RR102, RR999) with indicators confirming the presence of a building, such as positive assessed building value, market value, building area, gross building area, or the presence of a 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 also recommend that analysts double-check whether, in their study area, "RR" codes are based on legal zoning, or imply the presence of a building.
- **Vacant parcels** (parcels without any building) 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 (i) without building footprints, (ii) without a land use code indicating the presence of a building, and without a positive value for improvements in either the tax assessors' (iii) valuation or their (iv) fair market value estimates (e.g., Nolte 2020).
- 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 et al. 2016), and small errors can lead to small sample sizes or bias. Agricultural parcels can be found under a range of land use codes, including "AG" (agricultural), "VL" (vacant land) and "RR" (residential). Analysts filtering ZTRAX data for agricultural sales might therefore 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 the identification of irrigated cropland from governmental sources can help distinguish between irrigated and non-irrigated agricultural areas in sample selection and analysis.

3.5. Dealing with missing or mismeasured data for standard housing attributes

Relevance: Hedonic analyses need to 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 or through quasi-experimental matching techniques. This requires that these characteristics are reliably and consistently observed across the study region.

Problem: The availability of standard housing attributes in ZTRAX varies across counties, often clustered by state. Across all residential property records in ZAsmt, 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%) (Fig 5). Non-sensical zero values (e.g., zero rooms, zero living area) are not uncommon. When non-zero 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 near-complete availability of "lot size" data (Fig 5), 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 both across and within jurisdictions, presumably due to variability in practices between communities or individual assessors.

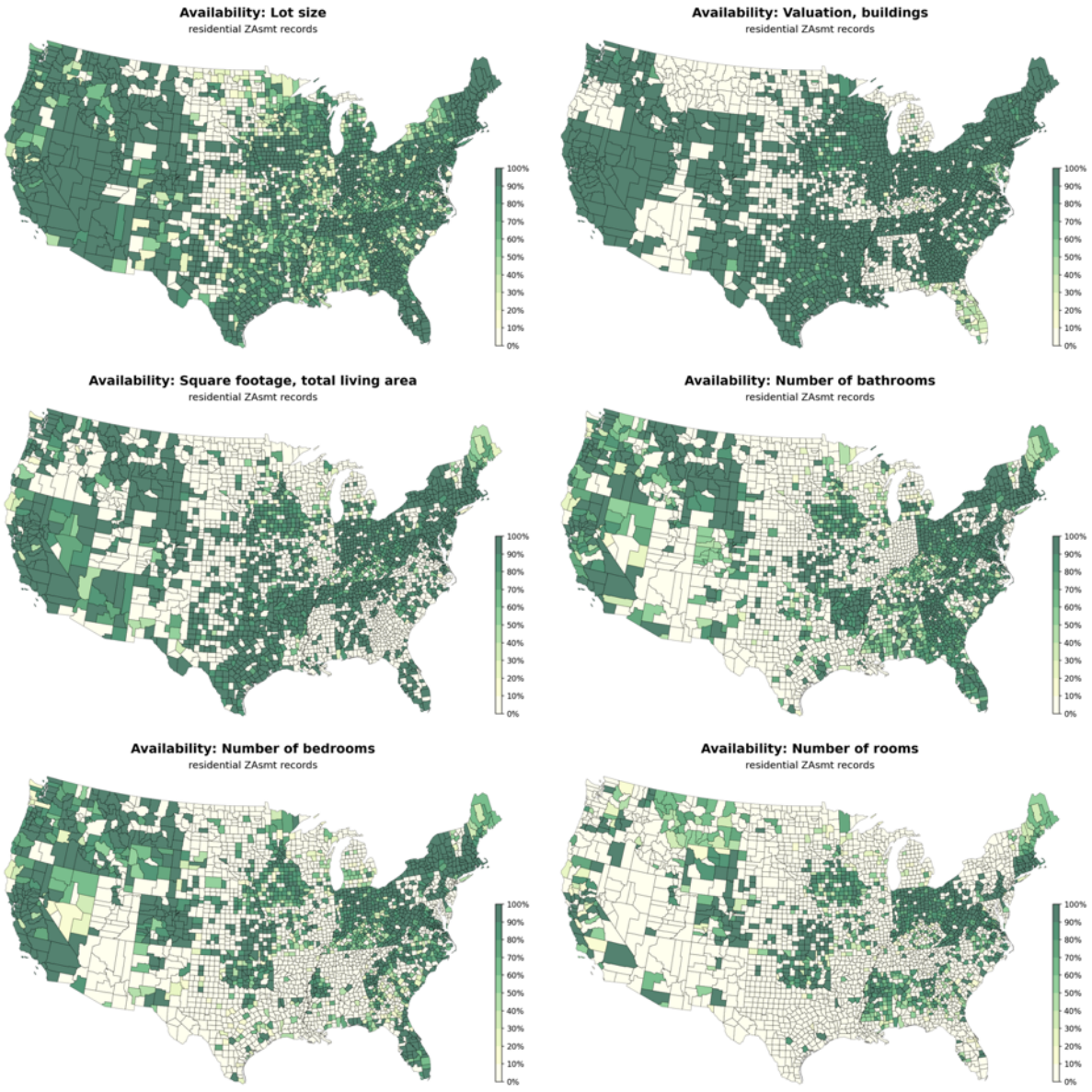


Fig 5: County-level availability of standard housing indicators for residential ('RR') parcels in ZTRAX tax assessment records (ZAsmt). Shown is the percentage of non-zero values for lot size (top left), building valuation (top right), square footage of living area (center left), number of bathrooms (center right), number of bedrooms (bottom left), and number of rooms (bottom right).

Missing data means that researchers face sample-selection bias issues, while unreliable data is a measurement error issue. Both are likely to 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 non-representative sample. A pairwise completeness for a subset of ZAsmt attributes indicates considerable heterogeneity of attribute completeness in ZAsmt (Fig 6): for example, the joint analysis of building square footage and land use type would cover a sample of 68% of the almost 150 million property records in ZAsmt. Analysts who instead account for missing observations in their empirical models need to consider how their choice of methods might bias their estimators and affect their geographic coverage. 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 contains new information (Moulton et al. 2018; Clarke & Freedman 2019; Fraenkel 2019; Gindelsky et al. 2019; Albouy et al. 2020). Analysts working with temporally varying characteristics from the historical ZAsmt database need to be aware of data gaps resulting from sub-county level updating cycles of the underlying assessor data that lead to incompleteness patterns which vary across space and time (Fig S16a). Methods to mitigate the resulting bias may include spatial aggregation (Fig S16b) or record-level time series interpolation (Fig S16c).

Potential Solutions: 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 & Trivedi (2005, chapter 26 and 27) 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 data inspection, robustness checks, and a full documentation of 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, etc.), or matched external validation data (e.g., lot sizes from parcel boundaries, jurisdiction, income, race, etc.). The appropriate data inspection approach should be guided by the analyst's research question and design. For instance, the pattern of missing/mismeasured data that cause bias are likely to be different between cross-sectional and panel or difference-and-difference models. 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 importantly, we recommend that analysts fully document their sampling procedure, including choices of inclusion vs. exclusion based on data availability and reliability, and the implications of that choice on the geographic coverage of findings.

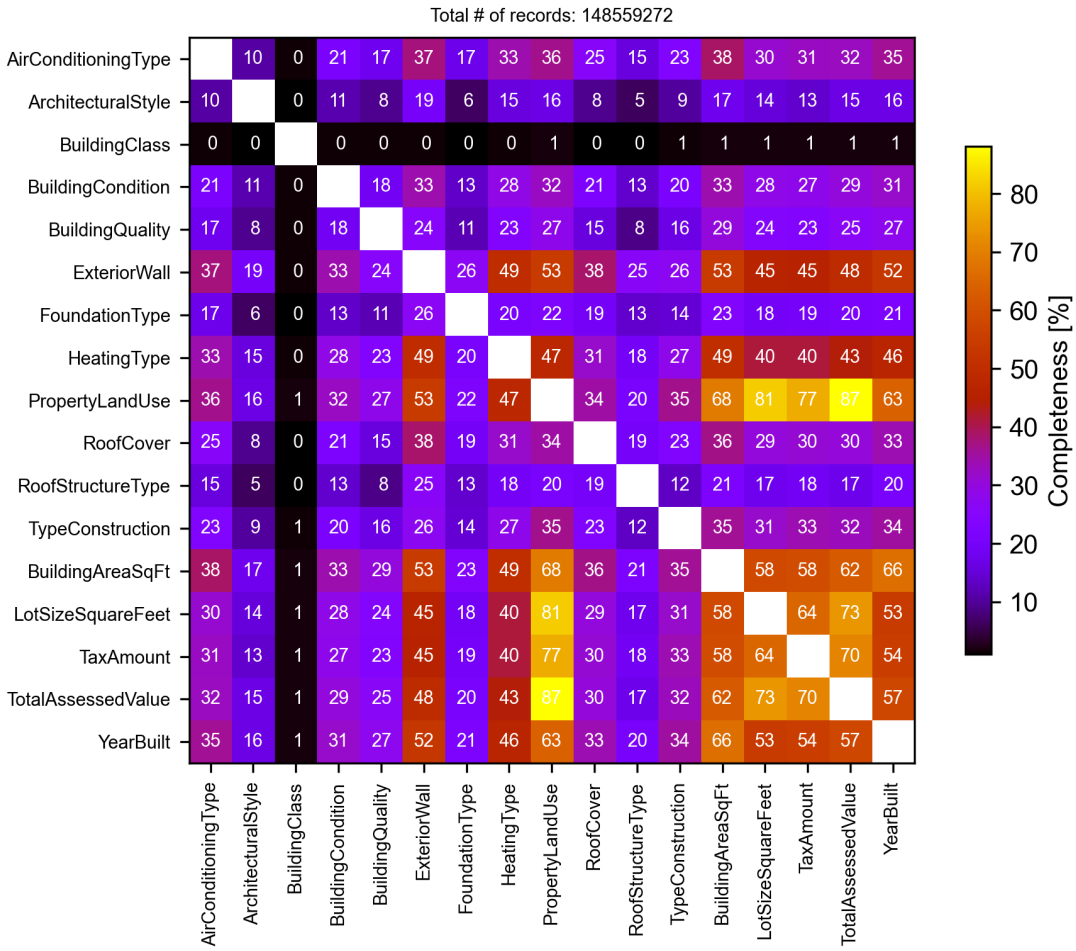


Fig 6: Mutual completeness of ZTRAX assessment records (measured in % of the total number of ZAsmt records) for selected building and property attributes, illustrating the variability of thematic data completeness across different attributes.

4. Conclusion

Nationwide property transaction and assessment data offer unprecedented opportunities for detailed empirical research into the dynamics of land ownership, land policy, and property valuation in the United States. For many scientific inquiries, particularly those in traditionally underfunded fields, ZTRAX might offer a "last chance", at least in the foreseeable future, to generate novel, large-scale insights at unprecedented detail and low cost. After conclusion of the ZTRAX program, analysts who compile data across counties and data sources, or those purchasing similar data services from third-party providers, will likely be confronted by many of the issues discussed here. Awareness of potential errors and biases in aggregated transaction and tax assessor data, full documentation of data processing and filtering choices, and the discussion of the potential effects of geographic omissions will enhance the quality, validity, generalizability, replicability, and interpretability of findings. We encourage journals editors and referees to have authors include detailed documentation of data processing choices in their final manuscript submissions. In the absence of official best practice standards, this document can serve as a non-exhaustive checklist of potential issues.

Supporting Information

Supplementary figures (Fig S1-S16) are included below.

Filtering tables and descriptions of deeds are provided as separate files.

The list of issues and potential solutions reported here is not exhaustive. Relevant updates will be posted to http://placeslab.org/ztrax.

Acknowledgements

We thank Zillow, Inc. for having made ZTRAX available free of charge for U.S. academic, non-profit, and governmental researchers. We thank participants of the 2021 PLACES webinar for useful feedback. Christoph Nolte, Ido Kushner, Adam Pollack, and Shelby Sundquist acknowledge support from the Department of Earth & Environment at Boston University, the Junior Faculty Fellows program of Boston University's Hariri Institute for Computing and Computational Science, and The Nature Conservancy. Johannes Uhl is funded, in part, by NSF's Humans, Disasters and the Built Environment (HDBE) program (Grant #1924670).

References

Albouy, D., Christensen, P. & Sarmiento-Barbieri, I. (2020). Unlocking amenities: Estimating public good complementarity. *J. Public Econ.*, 182, 104110.

Banzhaf, H.S. (2021). Difference-in-Differences Hedonics. *J. Polit. Econ.*, 129, 000–000.

Bernstein, A., Gustafson, M.T. & Lewis, R. (2019). Disaster on the horizon: The price effect of sea level rise. *J. financ. econ.*, 134, 253–272.

Bigelow, D., Borchers, A. & Hubbs, T. (2016). *U.S. Farmland Ownership, Tenure, and Transfer*. U.S. Department of Agriculture, Economic Research Service.

Bishop, K.C., Kuminoff, N. V., Banzhaf, H.S., Boyle, K.J., von Gravenitz, K., Pope, J.C., Smith, V.K. & Timmins, C.D. (2020). Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for Environmental Quality. *Rev. Environ. Econ. Policy*, 14, 260–281.

Brown, J.F., Tollerud, H.J., Barber, C.P., Zhou, Q., Dwyer, J.L., Vogelmann, J.E., Loveland, T.R., Woodcock, C.E., Stehman, S. V., Zhu, Z., Pengra, B.W., Smith, K., Horton, J.A., Xian, G., Auch, R.F., Sohl, T.L., Sayler, K.L., Gallant, A.L., Zelenak, D., Reker, R.R. & Rover, J. (2020). Lessons learned implementing an operational continuous United States national land change monitoring capability: The Land Change Monitoring, Assessment, and Projection (LCMAP) approach. *Remote Sens. Environ.*, 238, 111356.

Buchanan, M.K., Kulp, S., Cushing, L., Morello-Frosch, R., Nedwick, T. & Strauss, B. (2020). Sea level rise and coastal flooding threaten affordable housing. *Environ. Res. Lett.*, 15.

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

Clapp, C.M., Freeland, J., Ihlanfeldt, K. & Willardsen, K. (2018). The Fiscal Impacts of Alternative Land Uses. *Public Financ. Rev.*, 46, 850–878.

Clarke, W. & Freedman, M. (2019). The rise and effects of homeowners associations. *J. Urban Econ.*, 112, 1–15.

Connor, D.S., Gutmann, M.P., Cunningham, A.R., Clement, K.K. & Leyk, S. (2020). How

Entrenched Is the Spatial Structure of Inequality in Cities? Evidence from the Integration of Census and Housing Data for Denver from 1940 to 2016. *Ann. Am. Assoc. Geogr.*, 110, 1022–1039.
Fraenkel, R. (2019). Property Tax-Induced Mobility and Redistribution : Evidence from Mass Reappraisals ∗.
Gindelsky, M., Moulton, J. & Wentland, S. (2019). Valuing Housing Services in the Era of Big Data: A User Cost Approach Leveraging Zillow Microdata. *Int. Conf. Real Estate Stat.*
Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J.R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M. & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.*, 1–17.
Leonard, T., Yang, X. & Zhang, L. (2021). The impact of land use regulation across the conditional distribution of home prices: an application of quantile regression for group-level treatments. *Ann. Reg. Sci.*, 66, 655–676.
Leyk, S. & Uhl, J.H. (2018). Data descriptor: HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. *Sci. Data*, 5, 1–14.
Leyk, S., Uhl, J.H., Connor, D.S., Braswell, A.E., Mietkiewicz, N., Balch, J.K. & Gutmann, M. (2020). Two centuries of settlement and urban development in the United States. *Sci. Adv.*, 6, eaba2937.
Manson, S., Schroeder, J., Van Riper, D. & Ruggles, S. (2018). *IPUMS National Historical Geographical Information System: Version 13.0 [Database].* Minneapolis.
Microsoft. (2018). U.S. Building Footprints [WWW Document]. URL https://github.com/microsoft/USBuildingFootprints
Moore, M.R., Doubek, J.P., Xu, H. & Cardinale, B.J. (2020). Hedonic Price Estimates of Lake Water Quality: Valued Attribute, Instrumental Variables, and Ecological-Economic Benefits. *Ecol. Econ.*, 176, 106692.
Moulton, J.G., Sanders, N.J. & Wentland, S.A. (2018). Toxic Assets : How the Housing Market Responds to Environmental Information Shocks. *Work. Pap.*
Murfin, J. & Spiegel, M. (2020). Is the Risk of Sea Level Rise Capitalized in Residential Real Estate? *Rev. Financ. Stud.*, 33, 1217–1255.
Netusil, N.R., Moeltner, K. & Jarrad, M. (2019). Floodplain designation and property sale prices in an urban watershed. *Land use policy*, 88, 104112.
Nolte, C. (2020). High-resolution land value maps reveal underestimation of conservation costs in the United States. *Proc. Natl. Acad. Sci.*, 117, 29577–29583.
Olofsson, P., Holden, C.E., Bullock, E.L. & Woodcock, C.E. (2016). Time series analysis of satellite data reveals continuous deforestation of New England since the 1980s. *Environ. Res. Lett.*, 11, 064002.
Onda, K., Branham, J., BenDor, T.K., Kaza, N. & Salvesen, D. (2020). Does removal of federal subsidies discourage urban development? An evaluation of the US Coastal Barrier Resources Act. *PLoS One*, 15, e0233888.
Peng, L. & Zhang, L. (2019). House Prices and Systematic Risk: Evidence from Microdata. *Real Estate Econ.*, 1–24.
Pinter, N. & Rees, J.C. (2021). Assessing managed flood retreat and community relocation in the Midwest USA. *Nat. Hazards*, 107, 497–518.
Quesnel, K.J., Agrawal, S. & Ajami, N.K. (2020). Diverse paradigms of residential development

inform water use and drought-related conservation behavior. *Environ. Res. Lett.*, 15.
Shen, X., Liu, P., Qiu, Y. (Lucy), Patwardhan, A. & Vaishnav, P. (2021). Estimation of change in house sales prices in the United States after heat pump adoption. *Nat. Energy*, 6, 30–37.
U.S. Census Bureau. (2021). U.S. Census Geocoder [WWW Document]. URL https://geocoding.geo.census.gov/geocoder/
U.S. Geological Survey. (2019). Land Change Monitoring, Assessment, and Projection [WWW Document]. URL https://www.usgs.gov/land-resources/eros/lcmap
Uhl, J.H., Connor, D.S., Leyk, S. & Braswell, A.E. (2021a). A century of decoupling size and structure of urban spaces in the United States. *Commun. Earth Environ.*, 2.
Uhl, J.H., Leyk, S., McShane, C.M., Braswell, A.E., Connor, D.S. & Balk, D. (2021b). Fine-grained, spatiotemporal datasets measuring 200 years of land development in the United States. *Earth Syst. Sci. Data*, 13, 119–153.
Wentland, S.A., Ancona, Z.H., Bagstad, K.J., Boyd, J., Hass, J.L., Gindelsky, M. & Moulton, J.G. (2020). Accounting for land in the United States: Integrating physical land cover, land use, and monetary valuation. *Ecosyst. Serv.*, 46, 101178.
Zhang, L. & Leonard, T. (2021). External validity of hedonic price estimates: Heterogeneity in the price discount associated with having Black and Hispanic neighbors. *J. Reg. Sci.*, 61, 62–85.
Zillow. (2019). ZTRAX: Zillow Transaction and Assessor Dataset, 2019-Q4 [WWW Document]. URL http://www.zillow.com/ztrax
Zillow. (2021). ZTRAX: Frequently Asked Questions [WWW Document]. URL https://www.zillow.com/research/ztrax/ztrax-faqs/

Supplementary Material

Online Material (provided as separate files):

1. Table of fair-market value filters
2. Table of regular expressions to identify public sellers and buyers
3. Table of estimated geographic datums by county
4. Document of deed definitions

Supplementary Figures S1-S16 (see below)

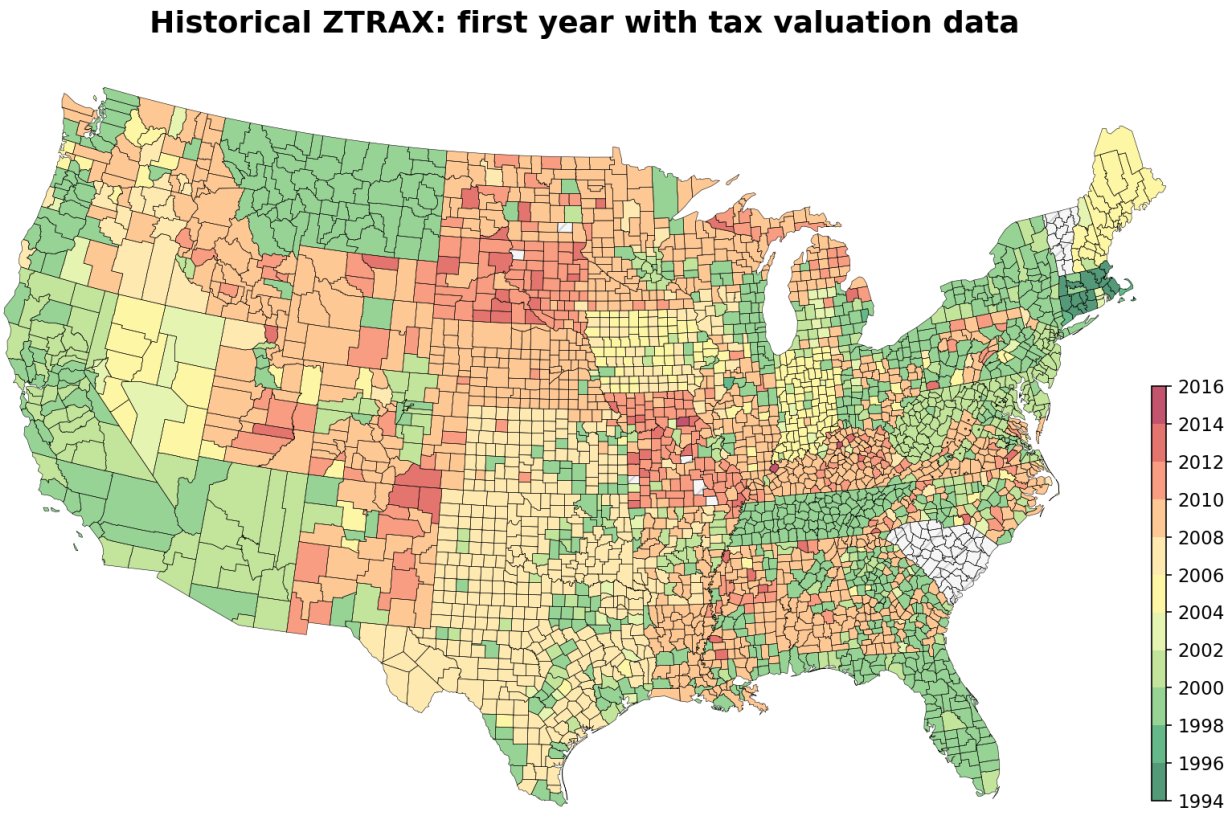


Fig S1: Earliest assessment year available in the historical ZAsmt table (table "utValue", column "AssessmentYear"). Grey color indicates an empty "utValue" table.

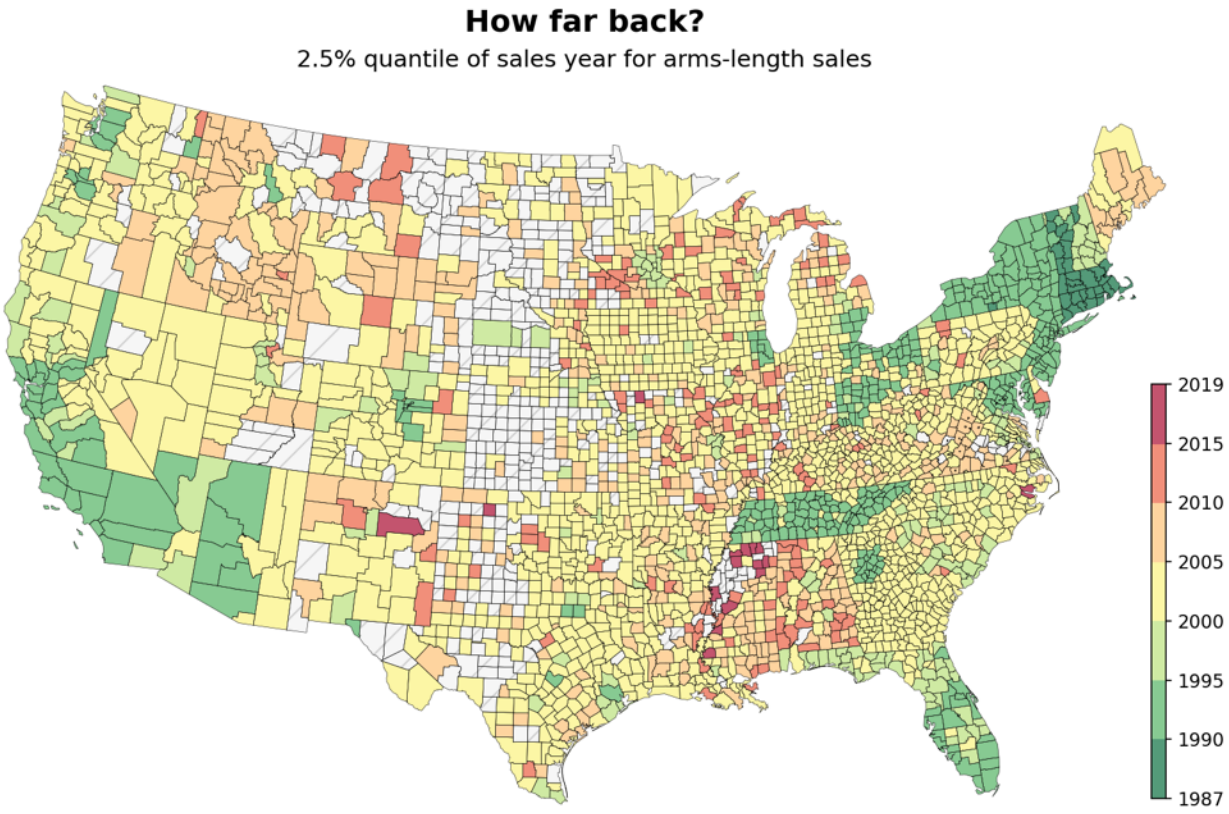


Fig S2: Temporal depth of ZTrans transaction records. Map shows 2.5% percentile of sale dates for arms-length sales in each county. Counties often contain a long and thin tail of early transactions, many of which might be based on erroneous date records. The 2.5% percentile therefore provides a more robust visual indicator for the beginning of the time period an analyst might reasonably choose as a cutoff for their analysis.

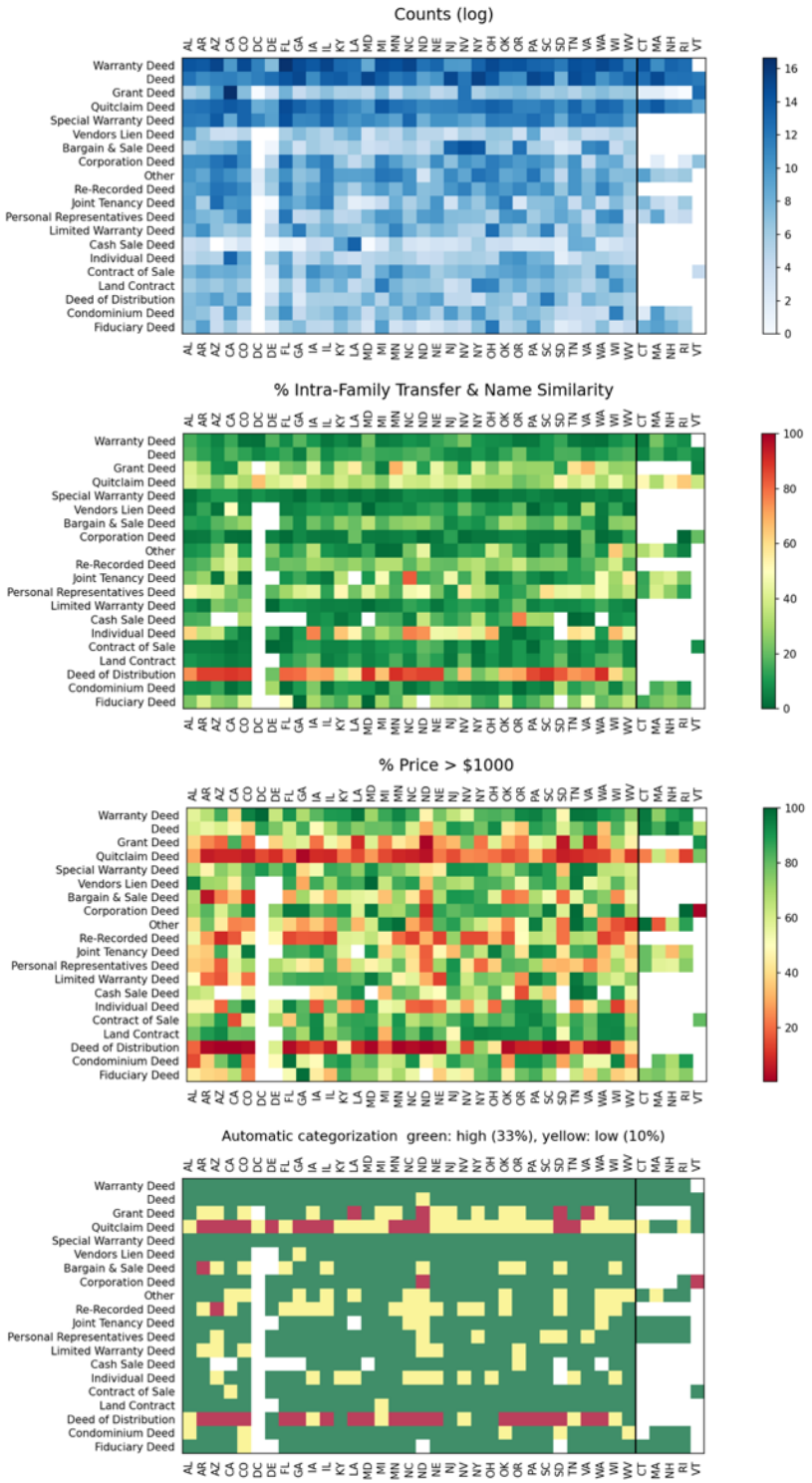


Fig S3: Illustration of our statistical approach to define confidence levels for combinations of document types and states, by example of the 20 most frequent deed types retained after the hierarchical exclusion of known document types (intra-family transfers, foreclosures). From top to bottom: frequency of document type; percentage of transactions identified as intra-family transfers (using both the intra-family transfer flag and similarity between buyer and seller names); percentage of transactions with prices larger than $1000; final categorization. States shown have price data in >20% of transaction records. New England states are shown separately to emphasize similarities in document type usage.

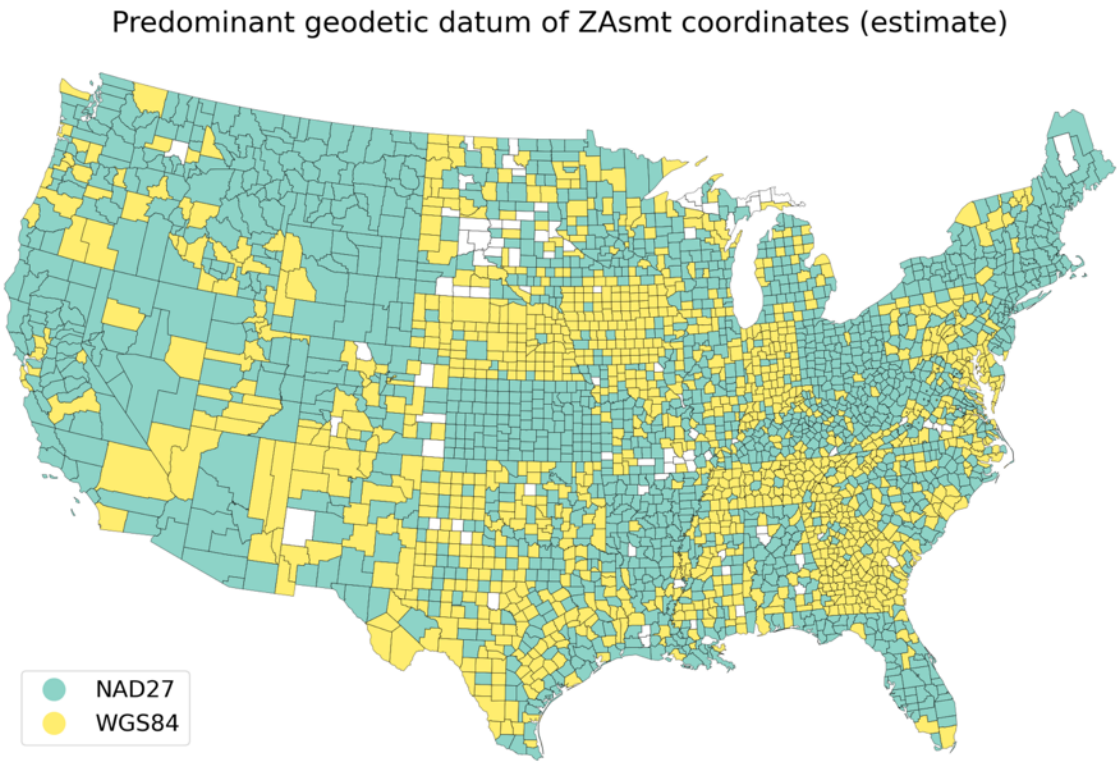
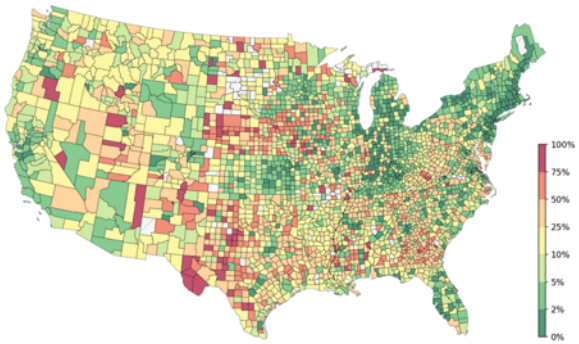
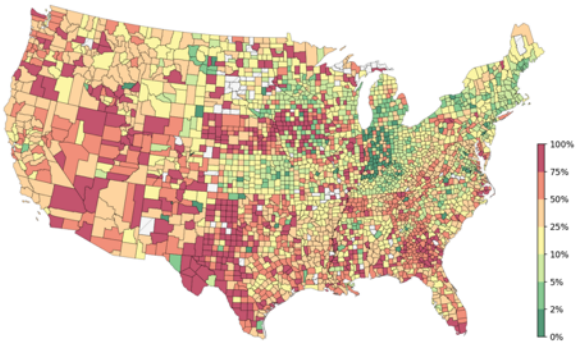


Fig S4: Estimated predominant geographic datum of ZTRAX coordinates. These estimates were obtained by projecting ZAsmt coordinates into a geographic projection using both the NAD27 (EPSG:4267) and WGS84 (EPSG:4326) datum and selecting the datum that produces the smallest median distance between ZAsmt coordinates and parcel centroids derived from geo-referenced digital parcel maps. Distance computation in the NAD83 Conus Albers projection (EPSG:5070).

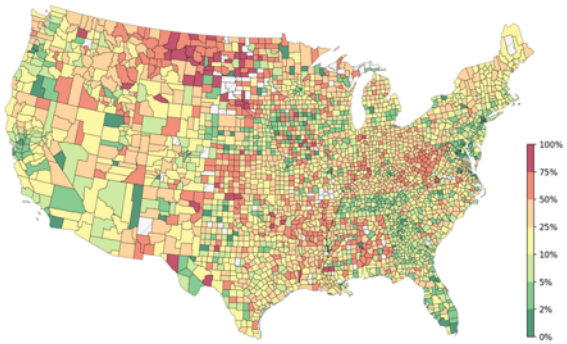
% missing coordinates, by count



% missing coordinates, by area



% duplicate coordinates, by count



% duplicate coordinates, by area

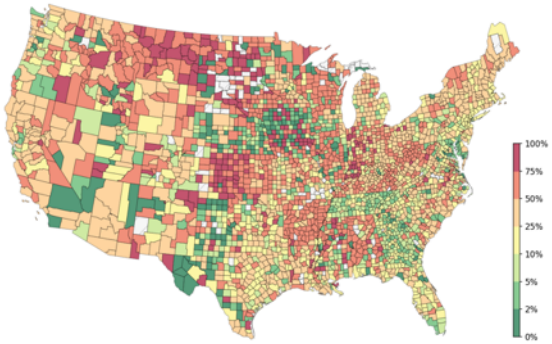


Fig S5: Prevalence of missing (top) and duplicate (bottom) coordinates in ZTRAX. Percentages are computed in reference to parcel count (left) and parcel area (right) for 131 million parcel polygons linked to ZAsmt records.

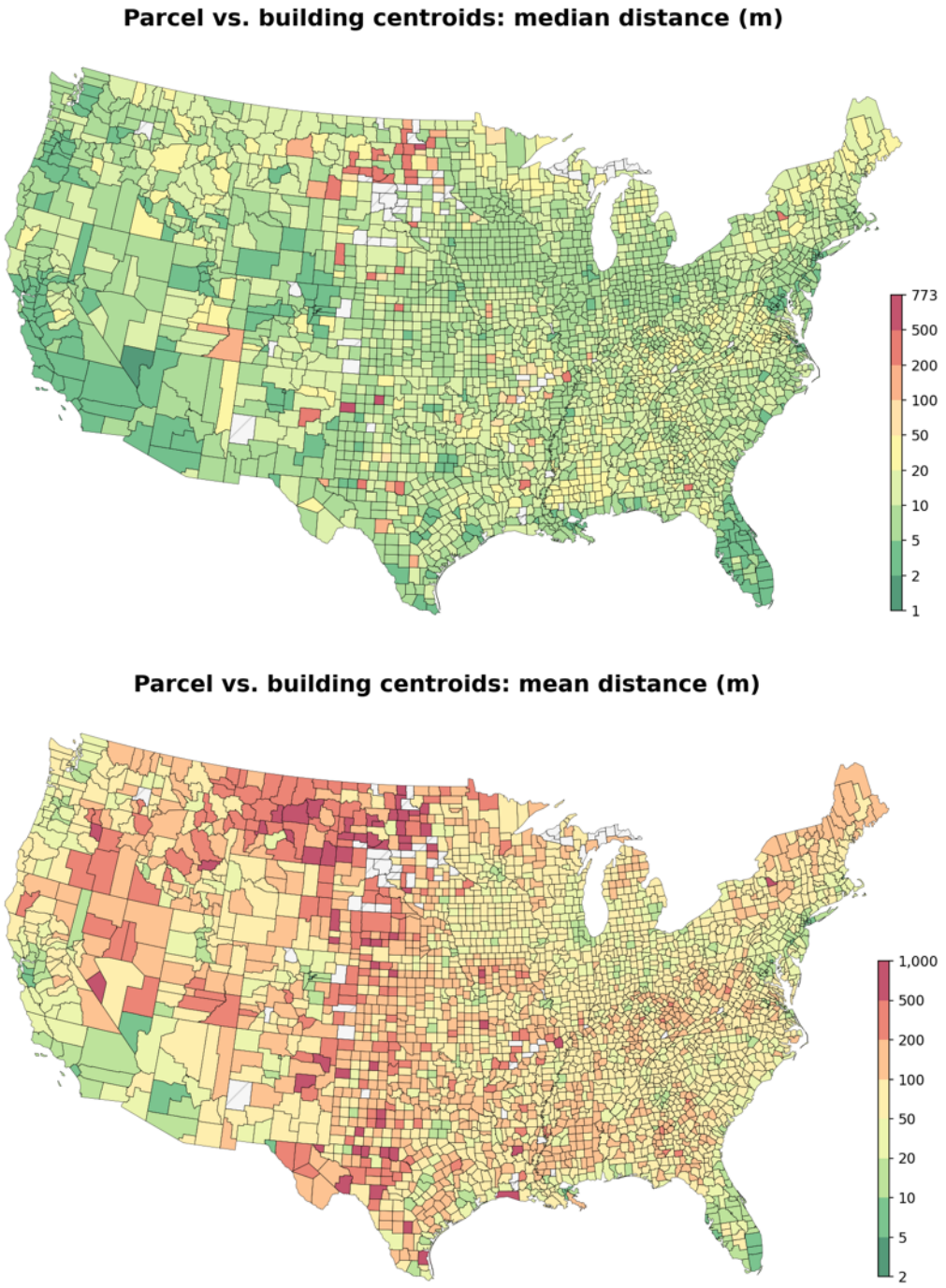


Fig S6: Median (top) and mean (bottom) distances between parcel centroids and building footprints for parcels with exactly one building footprint.

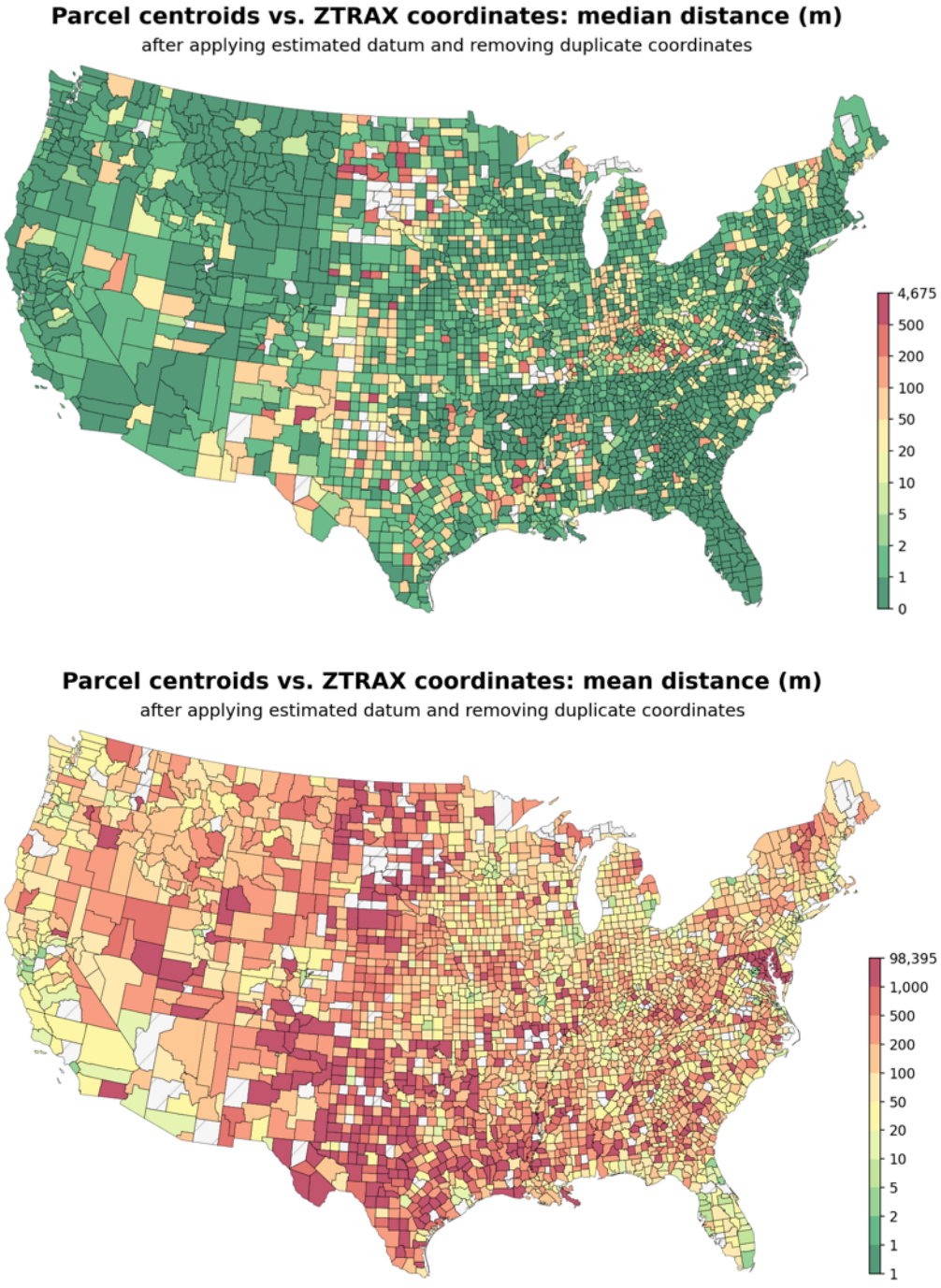


Fig S7: Median (top) and mean (bottom) distance between parcel centroids and ZAsmt coordinates for parcels uniquely linked to ZAsmt records. Distances are in meters and computed in EPSG:5070 projection.

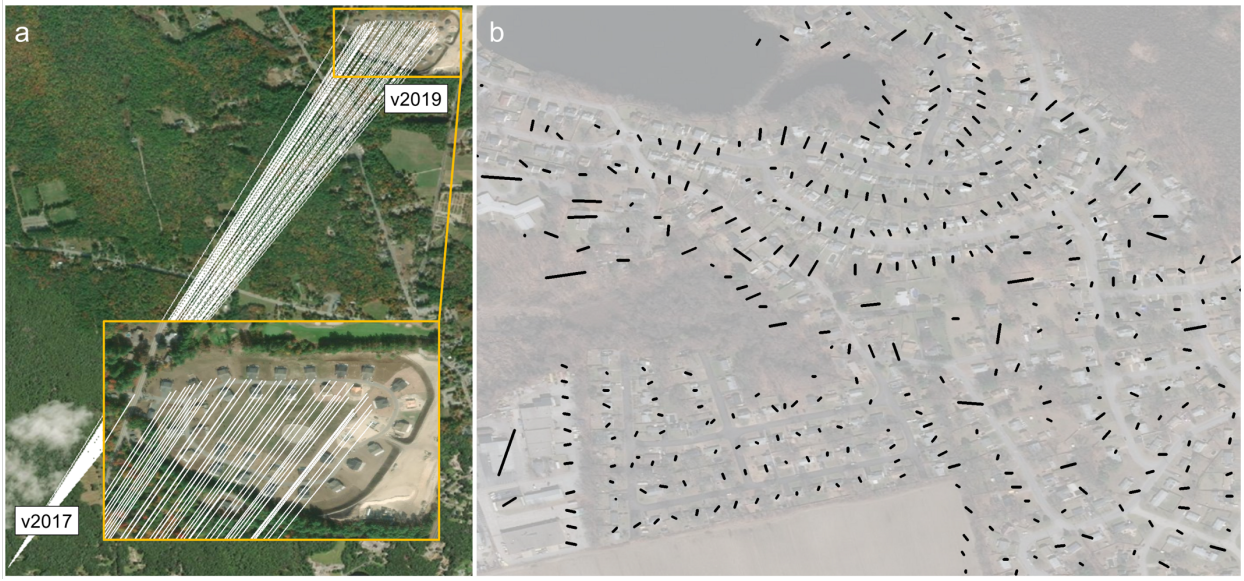


Fig S8: Illustrative example of observed changes in geo-coordinates between ZTRAX versions downloaded in 2017 and 2019. Panel (a) shows a new subdivision in Rhode Island. In 2017, all properties in the subdivision had identical geo-coordinates, likely derived from the ZIP code centroid. By 2019, each property had its own unique geo-coordinate, likely derived from street addresses. Panel (b) shows minor shifts in coordinates between 2017 and 2019 database version, likely resulting from a coordinate improvement based on building footprints. Imagery source: ESRI

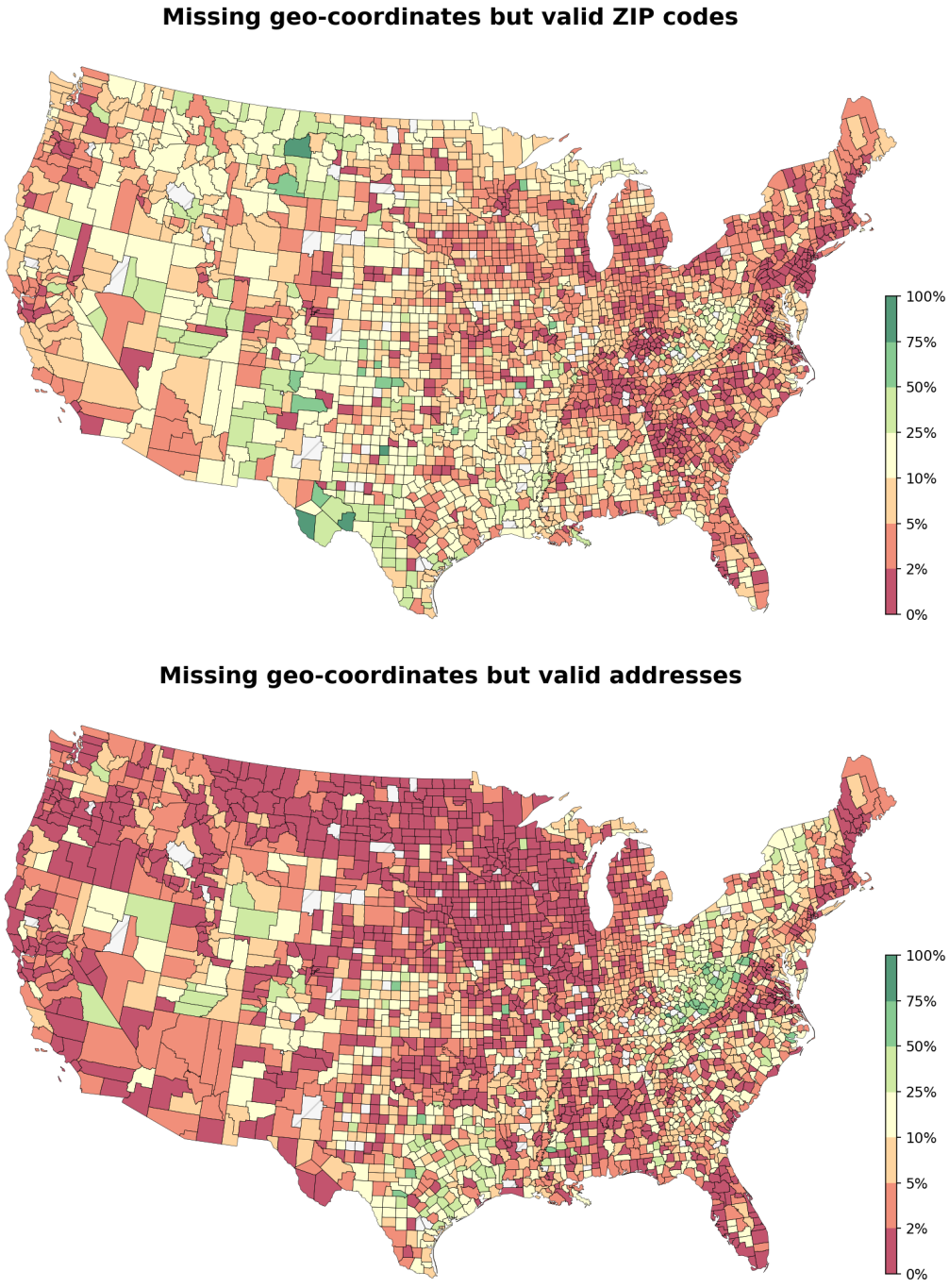


Fig S9: Percentage of records for potential improvement of geo-locations using the ZIP code and address information, shown here for a ZTRAX version from 2017.

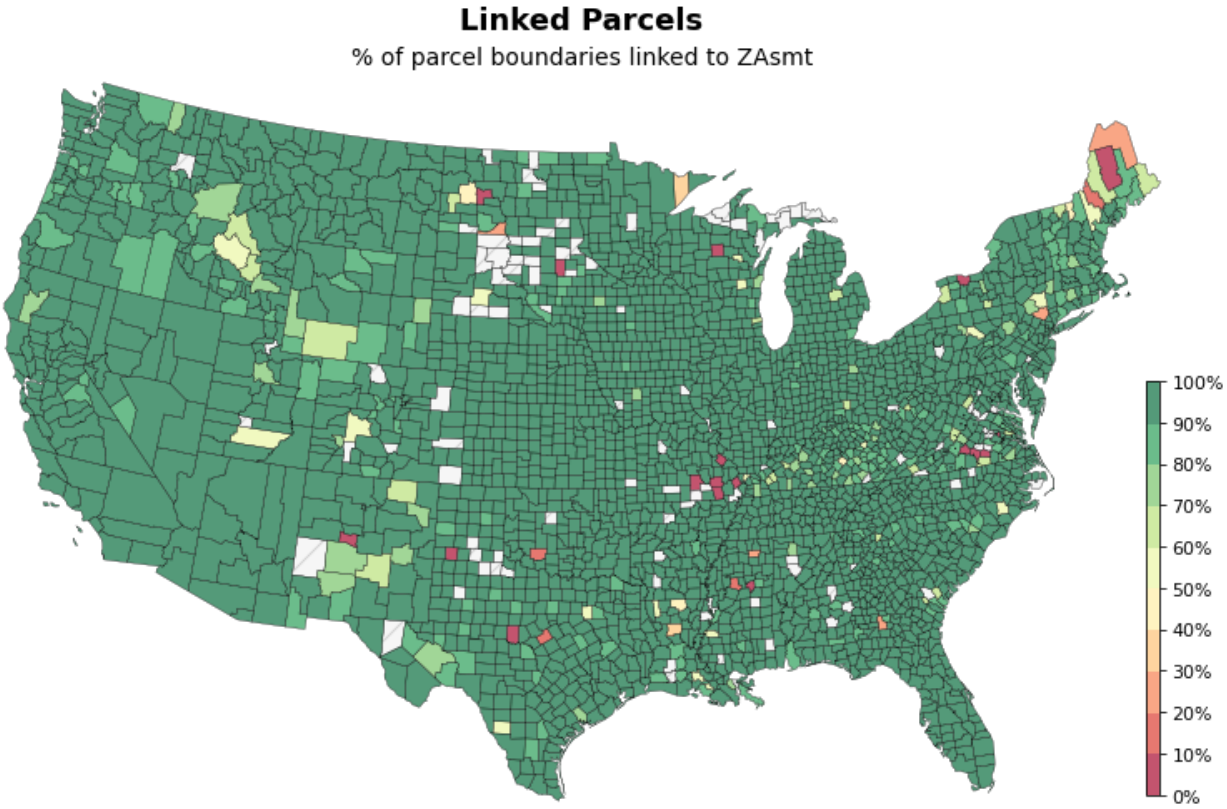


Fig S10: Percentage of parcels from digital parcel maps that PLACES lab members could successfully linked to ZAsmt records using string-pattern matching on assessor parcel numbers and tax account identifiers. Map shows status of PLACES data as of Jun 20, 2021. [Update maps after fixing recent losses in NY and Maine].

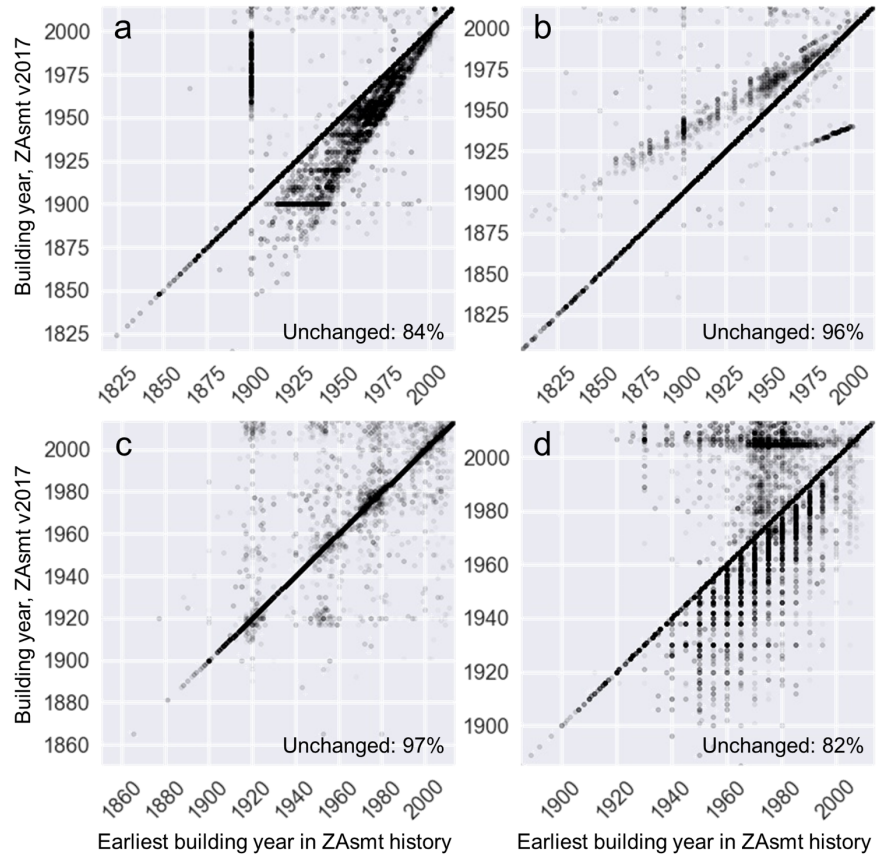


Fig S11: Differences in building year data for identical properties (matched by "ImportParcelID") between ZAsmt and historical ZAsmt versions in ZTRAX (both downloaded in 2017) for (a) Mercer County (Ohio), (b) Preble County (Ohio), (c) Natrona County (Wyoming) and (d) De Soto County (Florida). Points below the diagonal show records whose building years were backdated, i.e., building years in the historical ZAsmt version were replaced by an earlier year in ZAsmt. Note that these changes across database versions are relatively rare (82% - 97% in the examples shown).

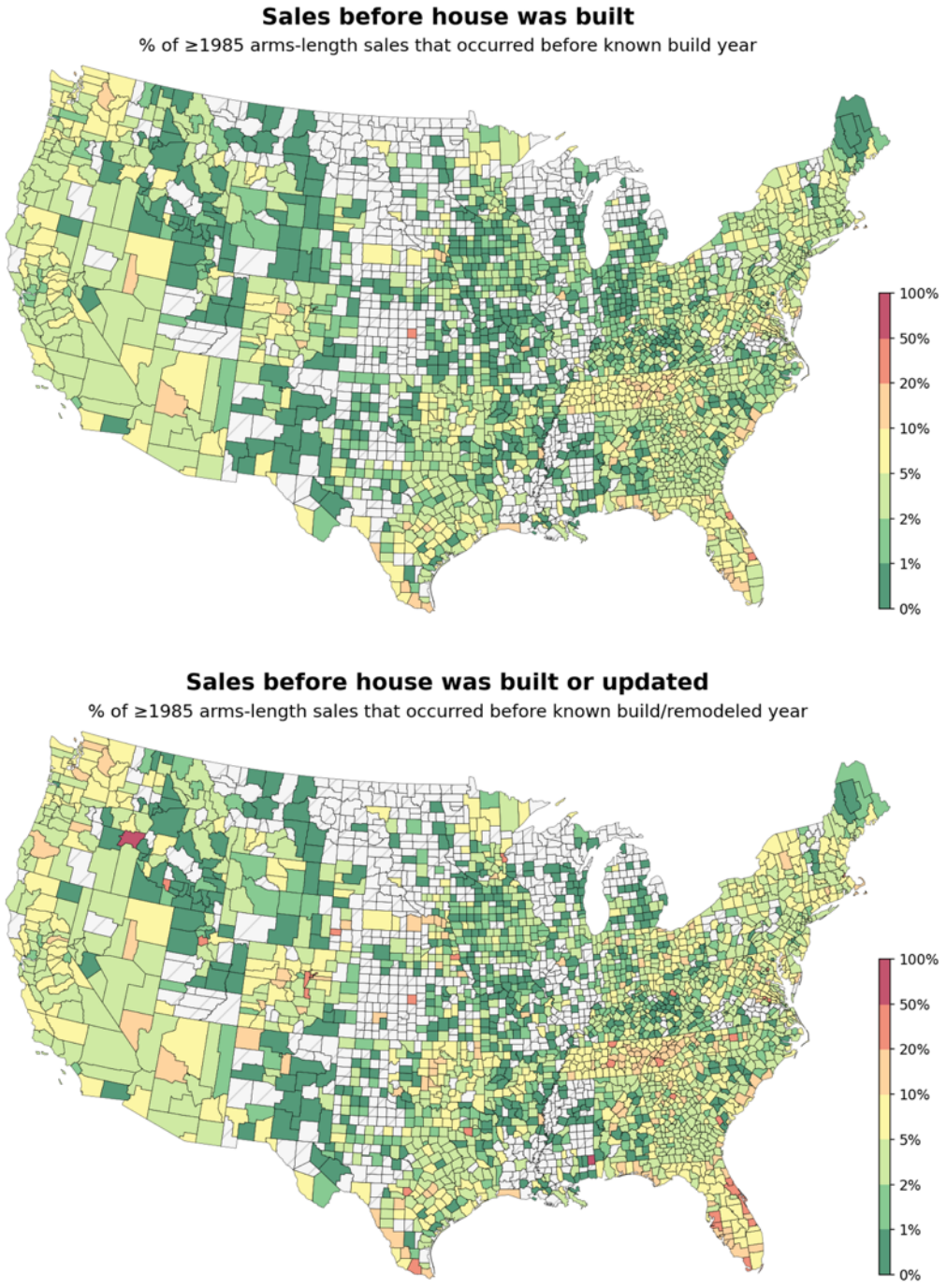


Fig S12: Percentage of sales observations after 1985 (parcel-linked, arms-length ZTrans records) that occurred before house was known to be built (top) or known to be either built or renovated (bottom).

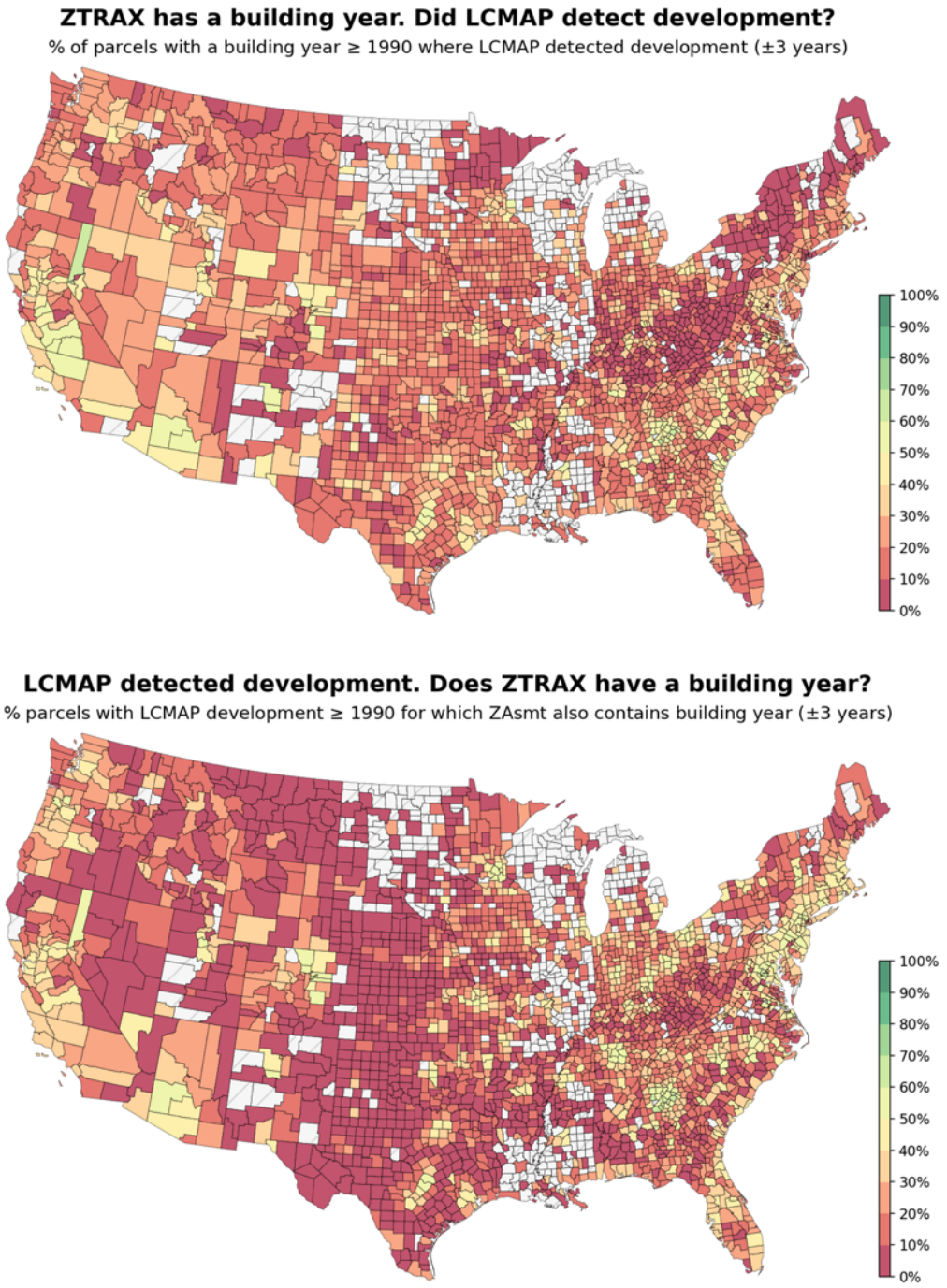


Fig S13: Discrepancies between building years reported in ZAsmt vs. most recent change of a pixel inside the corresponding digital parcel boundaries from a "non-developed" to a "developed" class in the LCMAP dataset (U.S. Geological Survey 2019).

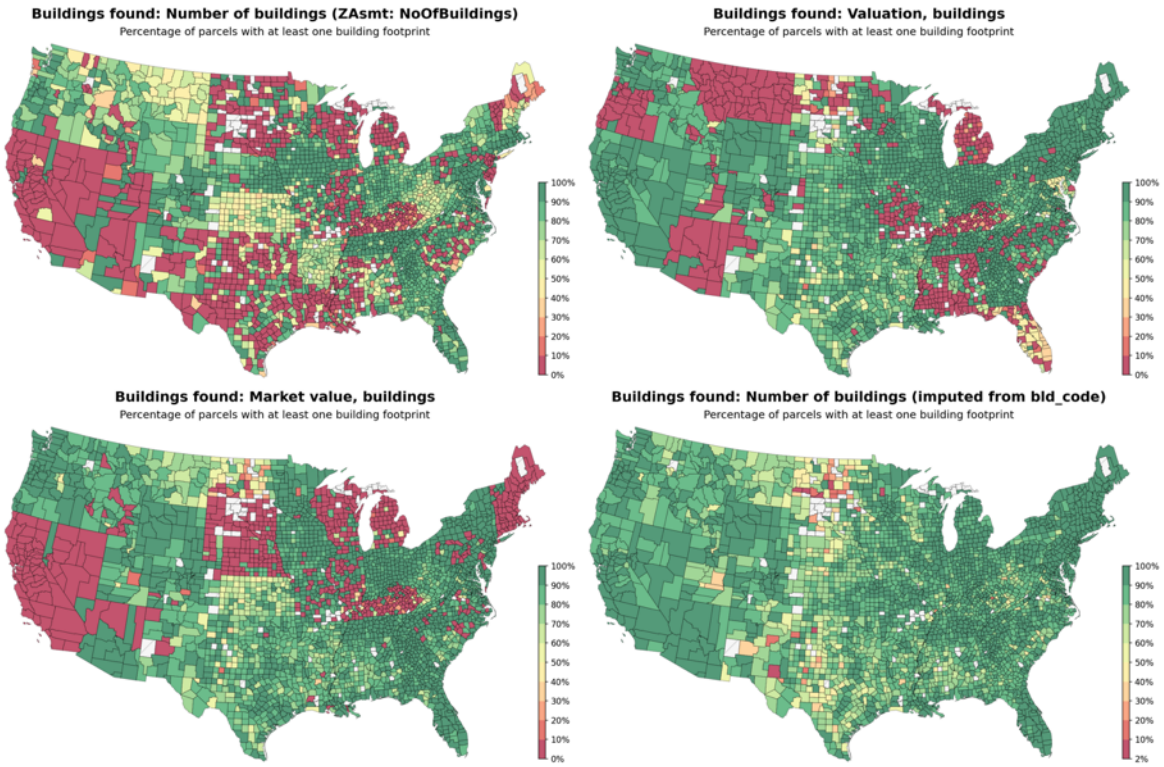


Fig S14: Percentage of parcels with at least one building footprint that would have been identified as "developed" based on the following four alternative indicators: Number of buildings (top left), positive building valuation (top right), positive building estimated market value (bottom left) and our own indicator of building presence derived heuristically from the standardized land use codes based on their text description (bottom right).

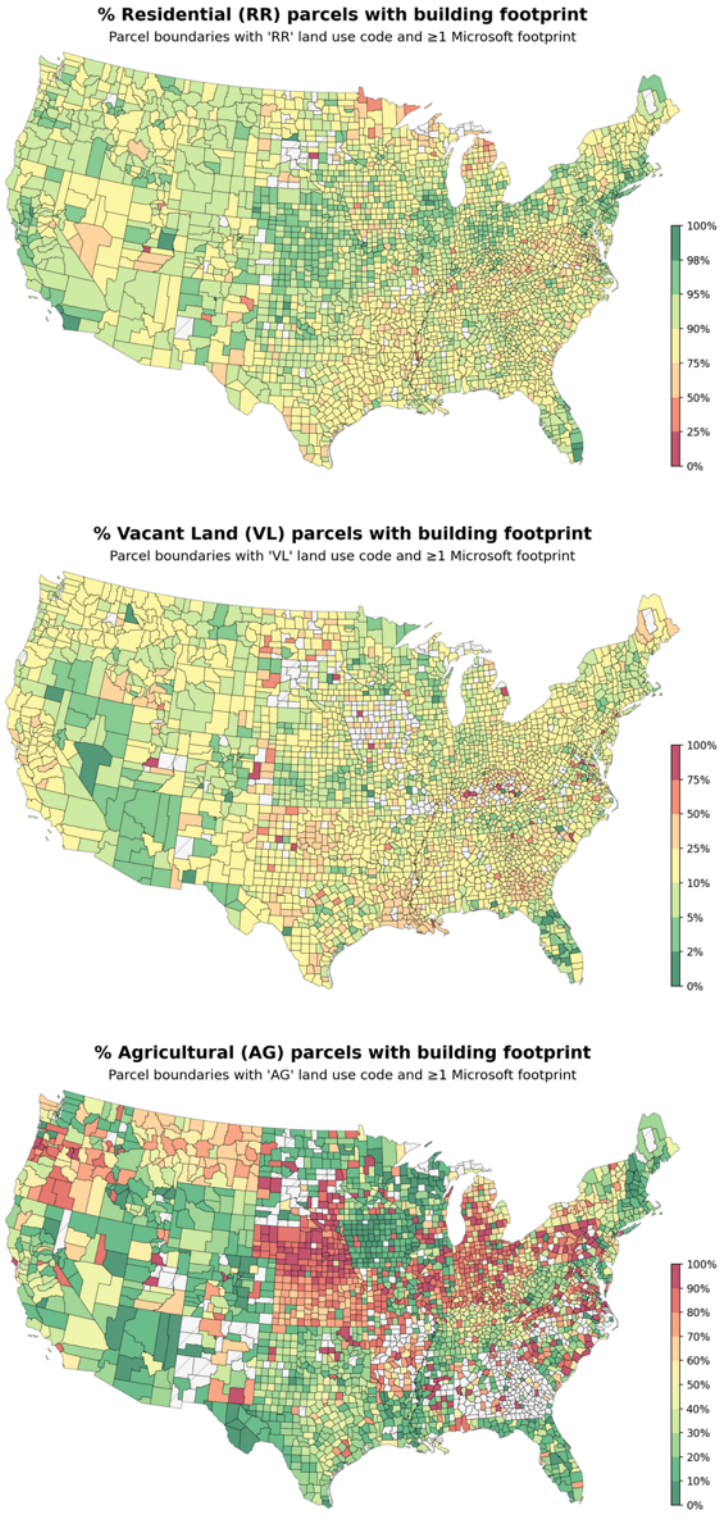


Fig S15: Percentage of building footprints found on "residential" (top), "vacant" (middle), and "agricultural" (bottom) parcels. The unique scale and direction of each map's color mapping serves to highlight concerns (absence of buildings on "residential" parcels, presence of buildings on "vacant" parcels) and heterogeneity in agricultural parcel definitions between neighboring states (e.g., Nebraska/Kansas vs. Iowa/Colorado)

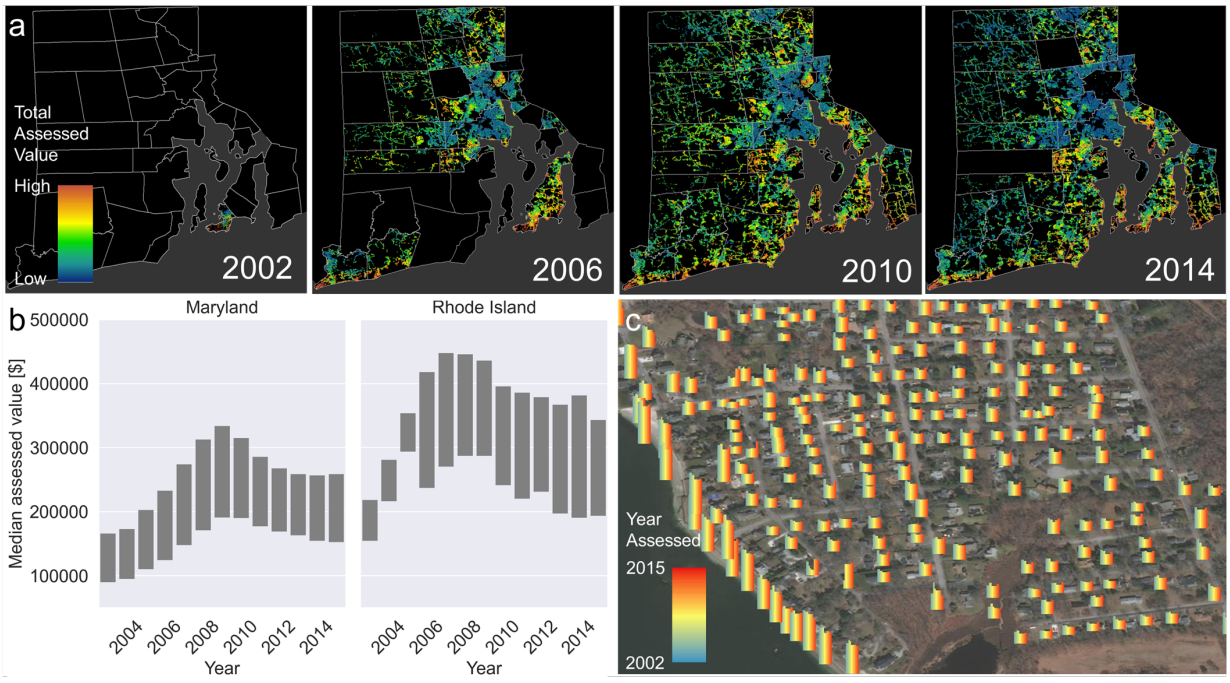


Fig S16: Illustration of data gaps in historical ZTRAX by example of the variable "total assessed value". Panel (a): Total assessed value of properties in Rhode Island, shown for different years based on the assessment year attribute, illustrating different updating cycles per town (white). Panels (b) and (c) illustrate exemplary strategies to mitigate effects of irregular updating cycles, such as (b) spatial aggregation (shown here: IQR of county-level median assessed values per state) and (c) interpolation of record-level time series for a subset of Rhode Island. The vertical dimension of the bar charts shows the total assessed value, and the horizontal component of the bar charts shows its variation over time for individual ZTRAX records at the respective locations, after applying forward-filling to the missing values in each record-level time series. Imagery source: ESRI