Data Screening and Synthesis to Develop High-Resolution Building Inventories for Regional Risk Assessment

Meredith Lochhead, Ádám Zsarnóczay, Gregory Deierlein
mlochhea@stanford.edu

Stanford University, Department of Civil and Environmental Engineering, Stanford, California, USA

Abstract

High-fidelity regional simulations of the impact of natural hazards on the built environment can be used to support disaster risk management and guide mitigation priorities. These assessments are underpinned by building inventories. Historically, building inventory development has primarily been driven by insurance companies and government agencies, typically targeting aggregate risk and impact measures. The growing feasibility of modeling impacts beyond aggregate loss and the growing interest in regional risk studies from a broad range of stakeholders are creating a need for detailed footprint-level building inventories. Current research studies often use varied data sources to describe the building inventory or use variable single-use methods to synthesize multiple datasets; however, few have assessed the impact of these inventory development decisions or the variability of the results. This study presents 1) a systematic framework for creating footprint-level building inventories through the synthesis of multiple data sources, 2) specific implementation methods for various data types, and 3) a quantitative evaluation of how inventory development decisions impact the resulting inventory makeup and quality for a case study city. Results show that the choice of input data sources and synthesis methods can lead to substantial differences in the resulting inventory. Furthermore, these differences are geospatially clustered and concentrate in certain types of buildings, which can lead to significant biases in the results. These findings underscore the need for more systematic and standardized approaches to building inventory development for regional natural hazard risk assessments.

Keywords: Building Inventory Development, Exposure Modeling, Regional Risk Assessment, Geospatial Data Synthesis, National Structure Inventory

1. Introduction

Regional modeling of the impact of natural hazards on the built environment plays an important role in quantifying risks and informing strategies to make communities more resilient to earthquakes, hurricanes, and other extreme disrupting events. While it is well established that natural hazard risk results from the intersection of hazard, exposure, and vulnerability [1, 2], progress in characterizing these factors has been uneven. Significant advances have been made in modeling the hazard and vulnerability components of risk, but the development of exposure models (building inventories) has received comparatively limited attention in academic research.

Notable efforts in building inventory development have historically been for the purpose of improving aggregate loss estimates, driven primarily by insurance companies, government agencies, and NGOs. Detailed building inventories were first developed in the Untied States for fire insurance purposes in the early 1900s, and since then, the insurance industry has remained the primary driver of large-scale building inventory development [3]. These resources, which often rely on proprietary data, are inaccessible to researchers and operate outside the peer-reviewed literature. In addition, several private companies sell building inventory datasets, including Cotality (formerly CoreLogic) [4], LightBox [5], ATTOM [6], and Regrid [7], among others. Beyond the insurance industry, government agencies have also developed broad building inventories for regional risk assessment. The most impactful among these is part of FEMA's Hazus software [8]. Hazus is an open methodology for risk assessment that includes a building inventory with aggregated information at the census block level for the entire United States. This dataset is a strong example of data curation and documentation, and it has been widely adopted by researchers. The quality and availability of the Hazus inventory has historically reduced the incentive for developing new, detailed building inventories in the US [9, 3]. Besides the Hazus team at FEMA, other government agencies have also created national-level resources through the synthesis of multiple data sources, including the National Structures Inventory (NSI) by the US Army Corps of Engineers [10], USA Structures by a collaboration between the Oak Ridge National

Laboratory, USGS, and FEMA's Response Geospatial Office [11, 12], and the global building inventory by USGS for their PAGER service [13]. These datasets have been widely adopted for emergency preparedness, flood insurance, rapid earthquake impact estimation, and other regional studies [14, 15, 16, 17, 18]. Finally, NGOs, particularly the Global Earthquake Model (GEM), have also made substantial contributions to global inventory development [19, 20, 2]. GEM and its collaborators have conducted many studies on global and country-scale inventories with various methodologies including disaggregating and inferring data using national census or survey data, synthesizing multiple available datasets, and employing both top-down and bottom-up methodologies [2, 21, 22].

Recent studies demonstrated the potential of high-resolution regional models at the level of a city, county, or urban area to support disaster risk management, inform policy decisions, and prioritize mitigation strategies [23, 24, 25, 26, 27, 28, 29]. With advances in computational power, increased open data availability, and the growing interest in metrics beyond aggregate loss, high-resolution exposure modeling at a regional scale is becoming more feasible for a broad range of stakeholders. Consequently, regional studies have been moving towards building-specific footprint-level (rather than aggregated) building inventories [30]. However, many of the aforementioned datasets, such as the Hazus 7.0 Inventory, GEM's Global Building Exposure Model, and the PAGER inventory, are only available at aggregated spatial resolutions, such as census tracts, and do not provide information at the individual building level [9, 2, 13]. As a result, although these datasets are designed for risk assessment and effective for determining aggregated loss and impact metrics, their aggregated nature limits their utility for building-level analysis by obscuring correlations between building-specific features and potentially introducing biases related to local site conditions [31]. The aformentioned inventories that do have building-specific information (e.g., NSI) often contain gaps with missing buildings or features, or have filled all gaps by statistically inferring or imputing data. This poses a challenge because data sources do not indicate which buildings or features are inferred, which makes it impossible to separate ground truth from statistically inferred fields. As a result, data quality may be inconsistent, and the reliability of available information is unclear. Due to these limitations, there is no obvious choice for experts who need a comprehensive building-level inventory dataset.

Across the many regional studies published in the literature, there is a wide variety of approaches for developing building-level inventories for risk assessment, the level of detail and specific methodology often depending on the scale of the analysis, available data, and specific objectives [32]. For example, if the area is reasonably small, some studies employ manual data collection, either in person or using street view imagery, to extract high-quality, location-specific building data [33, 21, 29, 34, 35, 36]. In slightly larger study areas where manual data collection is not feasible, many studies leverage publicly available administrative data such as tax assessor records, building permits, or other city-specific datasets to develop inventories, particularly as more municipalities transfer their records to GIS-based data repositories [3, 37, 28]. Others leverage portions of national datasets such as the census or national inventories like NSI due to their comprehensive coverage, availability, and ease of use [16, 17]. Complementing these approaches, advancements in remote sensing and imagery-based data collection, including satellite imagery, street-level imagery, and project-specific drone flights, have also allowed for the development of automated feature extraction via computer vision techniques [18]. For example, machine learning algorithms have been used to extract building footprints [38, 39], infer building features [40, 41], and classify building typologies [42, 43]. When detailed data is not otherwise available, researchers have worked with a variety of strategies to overcome data limitations, including simplifying assumptions, heuristic rulesets, and imputation to fill in gaps [44, 45, 26, 18]. Due to limitations in individual data sources and methods, several studies have developed building inventories by synthesizing multiple different data sources, thus producing a more comprehensive inventory [44, 34, 46]. Finally, some researchers have opted to develop synthetic inventories to bypass data availability challenges [47].

Despite the consensus on the importance of detailed inventories and the wide variety of available modeling approaches, many studies recognize the significant challenges in building inventory development. For example, manual data collection can be expensive and time consuming, and local data sources such as tax records can require large amounts of preprocessing and may still have gaps in important features. Furthermore, inventories that are based on inferred and imputed information, while comprehensive, may not have the local specificity or building-specific accuracy targeted in a study.

While studies in the literature utilize various inventory data sources and development methods, few have systematically evaluated how their decisions affect building inventory quality or regional risk assessment outcomes, and few have included uncertainty explicitly in their building inventories. There are two main challenges identified in the literature regarding building inventory quality, including 1) missing data, and 2) potentially incorrect data. Roohi et al. (2021) [48] explore the impact of missing data in the building inventory (as a proxy for inventory quality) and its implications for resilience metrics. Roohi et al. [48] find that reducing inventory accuracy, simulated by artificially removing data at random from a high-quality pre-existing inventory database, led to decreased reliability in estimating damage, functionality, and displacement. Sanderson and Cox (2023) [49] explore the second identified

inventory challenge, potentially incorrect data, by comparing a national-level inventory (NSI) and local tax parcel data for the small community of Seaside, Oregon. Sanderson and Cox [49] find that the two inventories show large differences at the parcel level, indicating possible inaccuracies within the NSI dataset, but also demonstrates that the two compare favorably in terms of structure value, year built, and plan area when aggregated at the block, block group, and tract level. When data sources disagree, as discussed in this study, one or both data sources have incorrect data for that building or feature. A separate case study in Luchon, France also explored the impact of inventory resolution (national versus local data) on the estimation of seismic damage and concluded that a major source of uncertainty in damage estimation is the "intrinsically difficult" inventory problem [50]. The two aforementioned challenges, missing data and potentially incorrect data, can only be evaluated directly if a high-quality "ground truth" dataset is available, which was the case in the aforementioned studies [48, 49, 50]. In the absence of a high quality "ground truth" dataset, others have sought to account for building inventory uncertainty directly. One study states that different exposure modeling hypotheses can lead to different building inventories and includes uncertainty by generating four different building inventories and conducting a risk assessment on each [51]. The results demonstrate the importance of incorporating detailed information and local knowledge in earthquake risk assessment because the estimated impact (e.g., expected loss or number of collapsed buildings) can vary by an order of magnitude. Others, while not propagating uncertainty explicitly, note both aleatory and epistemic uncertainty as important factors and current limitations of inventory development [52]. Despite these contributions and the importance of inventories as an input in risk modeling, the authors are not aware of any peer-reviewed studies that have explored best practices for synthesizing multiple inventory data sources into a single building inventory or assessed how the selected data sources and synthesis methods affect exposure accuracy. As a result, there is a lack of standardized workflows or guidance for researchers seeking to leverage multiple datasets to develop regional exposure models.

To address these gaps, this study presents 1) a systematic framework for creating a footprint-level building inventory through the synthesis of multiple data sources, and 2) a quantitative evaluation of how inventory development decisions impact the resulting inventory makeup and quality.

2. Inventory Development Concepts

In this study, we demonstrate that selecting different data sources and/or synthesis methods leads to substantially different inventories, which in turn could change the level and distribution of seismic risk estimates. Approaching inventory development in the same way the community approaches the hazard and vulnerability components of risk, with common best practices, established methods, and benchmarking, could allow for a more consistent and standardized framework for this process. In this section, for the sake of clarity, we discuss several specific concepts related to inventory development and define the corresponding terminology used throughout the study. Appendix A also contains a glossary of relevant terminology.

2.1. Inventory Resolution, Accuracy, and Fidelity

The methods and purpose of regional risk analysis are evolving beyond assessing aggregated damage and loss to focus on more localized and detailed metrics to better understand and plan for the impact of natural hazards [30]. The increasingly detailed questions posed about regional risk require higher-resolution inventories with specific building information. Inventory resolution consists of two components: spatial resolution and typology resolution. *Spatial resolution* refers to the geographic unit at which inventory information is provided. For example, data could be provided at the individual building level (higher spatial resolution) or aggregated at the census tract level (lower spatial resolution). Studies are using higher levels of *spatial resolution* by moving from aggregated census tract level analysis to footprint-level analysis [23]. Higher spatial resolution provides more geographic detail on the building inventory and allows for the incorporation of more local hazard effects [31]. On the other hand, *typology resolution* refers to the level of detail used to describe the building itself. For instance, a broad category like "light frame wood construction" represents a lower typology resolution, whereas a more specific description, such as "2-story wood house from 1962 with an elevated crawlspace," represents a higher typology resolution. Typology resolution is separate from the type of model used to assess performance (e.g., fragility function, SDOF oscillator, nonlinear finite element model) and focuses only on the building's physical description. Studies are also incorporating more detailed *typology resolution* by using more specific structural models, including surrogate models for regional assessment [53, 54]. These two metrics of resolution can change independently from each other.

Improved *accuracy* of inventories, or the degree to which the modeled inventory correctly represents the real-world building inventory, is another important consideration. Accuracy is independent of both spatial and typology resolution, and it does not automatically increase as resolution increases. Conversely, higher-resolution inventories

are challenging to represent accurately. For example, an inventory that is accurate at the aggregated census tract level (e.g., the Hazus inventory) does not automatically translate into high accuracy when disaggregated to the footprint level. Increasing the typology resolution presents even greater challenges as it demands the collection or inference of additional information (e.g., building code requirements in force when the building was constructed). While an inventory may correctly label a building footprint as a single-family house, determining accurate information on the year built, presence of a crawlspace, or other relevant features is a difficult task. If such information is not available and they are inferred, the accuracy of the inventory is typically reduced.

In this paper, the *fidelity* of a building inventory refers to the extent to which the inventory accurately and comprehensively represents the real building stock. Creating higher-fidelity inventories involves increasing the spatial and/or typology resolution, while maintaining at least the accuracy of the aggregated inventory. Fidelity represents a balance between the benefits of higher resolution and the potential loss of accuracy. If increasing spatial and/or typology resolution can only be done by disaggregating data through random sampling, it would increase the resolution but not the fidelity because no additional insight is being added through this process. For example, while naive disaggregation of a Hazus inventory into building footprints will increase spatial resolution, the fidelity of the inventory is unchanged if no additional information is incorporated. Fidelity only increases when additional information is used to refine the resolution, such as by incorporating information from independent national or local datasets.

When embarking on a high-fidelity regional assessment, it is important to consider the purpose of the assessment and whether the additional effort associated with developing high-resolution inventories is justified. For example, a high-resolution inventory may be unnecessary if a study focuses only on aggregate losses, but it is warranted if the assessment aims to quantify the risk to specific seismically vulnerable structures or buildings with important functions. In the latter example, detailed, specific structural models can be deployed more accurately if the inventory has sufficient building features to appropriately inform model selection. Increasing spatial resolution can also better capture realistic correlations in building features, which are lost or idealized when inventories are aggregated to the census block or tract level. Regardless of the selected spatial and/or typology resolution, it is important to maintain sufficient inventory accuracy to obtain reliable results from regional assessments. Conclusions should not be derived from data at a spatial or typology resolution that is higher than the level at which the inventory is considered sufficiently accurate.

2.2. Data Source Classification

This study presents a broadly applicable, general framework for building-level inventory development. The way in which the framework is implemented in a given case study hinges on what types of data are available, and thus, to inform and develop this framework, we first reviewed available resources and datasets that are relevant for inventory development. This section provides a brief overview of commonly used data types and calls attention to widely-recognized caveats about various public data sources. More specific commentary on individual data sources is included in Appendix D. The discussion primarily focuses on data sources within the United States, but similar concepts can be applied to regions outside the US. While there may be other important data sources, this list is made up of the sources that the authors are aware of and think are most relevant for building inventory development.

As mentioned in the previous section, there are several building inventory datasets explicitly designed for natural hazard risk assessment. The focus of this study is on building-specific inventories, and thus, this section does not go in depth on the Hazus, GEM, or PAGER inventories, which are spatially aggregated. At the building-specific level, there are several main nationally-available data sources. The National Structure Inventory (NSI) is a point-based dataset from the U.S. Army Corps of Engineers that describes building features for all structures across the United States [10]. It has been used in regional risk assessments [16, 49] and partially informs the Hazus inventory [55]. Similarly, USA Structures, a footprint-based dataset, includes information on occupancy class and building size, and it is used by FEMA for flood insurance mitigation [11]. Additionally, the Homeland Infrastructure Foundation-Level Data (HIFLD) includes occupancy-specific datasets for critical infrastructure across the United States [56]. These national datasets are easy to adopt due to their standardized formats, broad geographic coverage, and relatively complete data (typically few to no gaps or holes in the data). However, one challenges in comprehensive building-level datasets at the national level is that some information may be imputed or inferred, making it difficult or impossible to distinguish between what is true data and what is statistically sampled, potentially limiting local specificity and building-level accuracy.

At the local level, more granular datasets are often available from county or municipal sources. These can include local property assessor data (tax records), address points, zoning data, building permits, and other records. However, such data is not always available to the general public, and since there is no national standard, every

municipality uses its own unique schema when preparing such data. There are typically missing or inconsistently defined fields as well. Furthermore, these datasets are not designed for natural hazard risk assessment and it takes significant pre-processing effort to convert them into a usable format. Nevertheless, local datasets offer greater accuracy and detail at the building level, as they are based on ground truth either collected by official representatives of the city (in the case of tax data) or submitted by the owners of the buildings (in the case of building permits).

There are also several datasets with global coverage of building footprints, which can be created using computer vision algorithms (e.g., Microsoft Building Footprints [57]), can be community-developed (e.g., OpenStreetMap [58]), or developed by synthesizing multiple individual sources (e.g., Overture Buildings [59]). Most global building footprint datasets were not originally developed for natural hazard risk assessment, and while they are valuable for identifying buildings and supporting data synthesis, they generally lack detailed building feature information.

As mentioned in the introduction, there are also proprietary datasets describing the building inventory as well, often compiled by property data, real estate, or analytics companies. While some sell building inventory data (Cotality [4], LightBox [5], ATTOM [6], and Regrid [7], for example), others do not. For instance, Zillow [60] currently limits access to aggregated data at the zip code or neighborhood level, which restricts its usefulness for building-specific analysis. While these datasets can be beneficial for understanding the building inventory, the methodology proposed here aims to be relevant for other researchers without the need for costly private data. Thus, the proprietary inventory datasets are considered out of scope for this study, and the remainder of the discussion focuses on publicly available data.

Finally, there is an additional class of visual data that can be used to develop building inventories, including satellite imagery [61, 62], street-level imagery [63, 64], LiDAR data, drone data, and more. While increasingly available, these data types often require advanced computer vision or data inference techniques which can require significant computational data processing. Furthermore, there are often limitations due to poor image quality or occlusions. In this study, we aim to introduce a flexible, usable framework for creating a footprint-level building inventory that does not require significant image processing. For these reasons, image-based datasets are considered beyond the scope of this study. However, the inventory synthesis framework is defined such that inventory data derived from computer vision methods would be easy to integrate.

For the sake of clarity, several terms are adopted in this study to classify these inventory data sources. First, we refer to the spatial representation of each data source as its *source geometry*, which describes both the geometry (i.e., point vs polygon) and the scope of what the data describes (i.e., single building, multiple buildings, single unit). In this study, we classify data sources based on the following source geometry types, which are later used to inform how each source is synthesized and combined with other data.

- *Single-building polygons*: a polygon describing a single building, such as a building footprint database
- *Single-building points*: a point describing a single building, such as points located at the building centroid
- *Multi-building polygon*s: a single polygon describing multiple buildings, such as a census block polygon
- *Multi-building points*: a single point describing multiple buildings, such as a single point representing a whole mobile home park or university campus
- *Single-unit polygons*: a polygon describing a single unit, such as tax parcels denoting an individual condominium unit within a larger building
- *Single-unit points*: a point describing a single unit, such as point data located at every individual unit address
- *Mixed-type polygons*: Data that contains multiple types of polygon source geometries, such as tax parcels; tax parcels can describe a single building (single-family home), multiple buildings (apartment complex with a single owner), or a single unit (single condominium within a larger building)
- *Mixed-type points*: Data that contains multiple types of point source geometries, similar to mixed-type polygons
- *Non-spatial data*: Data that is not geospatially located, and thus does not have a source geometry

Second, *building features* refer to all characteristics of a building other than its source geometry. These include structural attributes (e.g., structure type, number of stories), use-related information (e.g., occupancy type), and socio-economic characteristics (e.g., number of residents, income, and ownership status). Similarly, *building feature values* refer to the specific entries used by each source to represent a building feature (e.g., for a 'number of stories' feature, possible values include 1 and 2).

3. Inventory Synthesis Framework

An important and well-known challenge in synthesizing multiple data sources is that each data source represents buildings differently in space (i.e., they have different source geometries). Thus, for data to be synthesized, all sources must be reconciled by co-locating data into a common *baseline geometry*, which is used to spatially describe the final inventory. The process of co-locating data into a single baseline geometry is not trivial. Inventory data sources are often geospatially approximate, and there is usually no one-to-one correspondence between different sources. Figures 1a through 1c illustrate such discrepancies between building centroids and footprints. Figure 1a shows points that fall just outside of the corresponding footprints (which is problematic for spatial joins based on overlap and intersection), cases where multiple points fall within a single footprint, and cases where footprints have no associated points. Figure 1b shows an example where points don't clearly relate to any footprint, including where points are located along transit lines (see expanded view in Figure 1c).

Similar issues arise between building footprints and parcel boundaries. In Figure 1d, one footprint comprises multiple parcels, which can occur in multi-owner buildings such as condominiums. Conversely, Figure 1e shows multiple footprints within a single parcel, such as an apartment complex where the parcel has a single owner. Figure 1f illustrates a case where the relationship between parcels and footprints is generally unclear. Even when parcels and footprints do map one-to-one, slight differences in geometry can make it difficult to assign those relationships systematically. Traditional spatial merging techniques often fall short in these situations. Carefully attributing data to a common baseline geometry can ensure that no data is inadvertently dropped and can help produce more robust results.

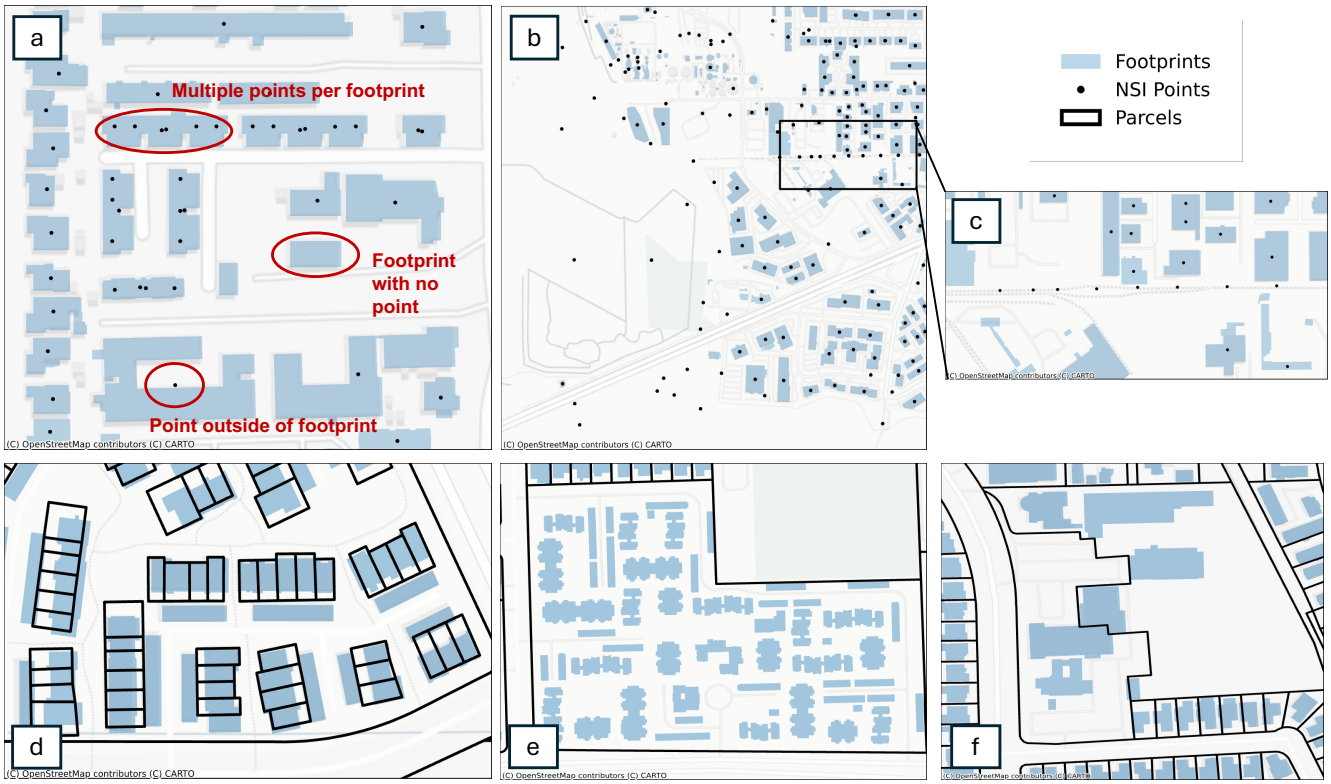

Figure 1: Examples of approximate geometries and lack of 1:1 mapping across source geometries for points and footprints (a-c) and parcels and footprints (d-f).

Based on these challenges, it is difficult to synthesize multiple data sources into a single, robust building inventory; even then, there may be gaps, inconsistencies, or other issues in the data. This section outlines several key steps for developing a building inventory, then presents the framework to facilitate the synthesis of multiple data sources. This framework was originally developed for a single-building baseline geometry, and for the sake of clarity, the descriptions and figures below assume footprints as the baseline geometry.

Furthermore, although this framework focuses on creating deterministic inventories by synthesizing multiple data sources, it is also important to acknowledge uncertainty in the building inventory. Unlike other sources of uncertainty in regional risk assessments, inventory uncertainty is epistemic, stemming from insufficient data about

the physical world, not inherent randomness. Although quantifying and propagating this uncertainty is not in the scope of this study, we discuss sources of inventory uncertainty and how this framework could support future uncertainty analysis.

The key steps in developing a building inventory are as follows:

1. **Select an appropriate resolution:** Based on the goals of the study, select the appropriate spatial and typology resolution needed in the inventory. As a practical matter, it is recommended to use the lowest (simplest) spatial and typology resolution necessary to address these goals.
2. **Select data sources:** Based on the chosen spatial and typology resolution, select data sources that collectively provide the required building features. Required building features are defined by the typology resolution and typically represent the minimum information necessary to enable downstream analysis (i.e., selection of a structural vulnerability model, evaluation of impact). It is critical to evaluate each data source for location availability (whether it covers the area of interest), comprehensiveness (what buildings it represents), accessibility (whether the data is readily available to use), and level of aggregation (building level vs. census tract, etc.). It is also helpful to consider the characteristics, gaps, and inaccuracies of the data source. One way to evaluate these is to consider data provenance, including the creator, host, purpose, creation date, last update, and the point in time the data most accurately reflects. The database's original purpose likely determines what is prioritized during its development. For example, national datasets like NSI are designed to be comprehensive, and thus contain very few gaps. This necessitates filling gaps in available data through disaggregation and inference based on approximate assumptions, which could lead to inaccuracies at the building scale. Conversely, local tax databases tend to be more accurate but lack national standardization and often omit detailed structural features, as they were not developed for risk assessment purposes.
3. **Select baseline geometry:** Select a baseline geometry appropriate to the study's spatial resolution. For example, if the resolution targets building-level analysis, the selected baseline geometry should be from a single-building source. This study recommends building footprints as the baseline geometry for several reasons. First, footprints have physical meaning. Unlike tax parcels, which may or may not correspond to a single building, footprints represent single buildings with consistent features. Second, as polygon geometries, they enable easier linking between datasets, especially for large buildings where point data may be scattered. Third, specifying inventory at the individual building level preserves correlations in building features that would be lost through aggregation. When selecting a baseline geometry source, it's important to screen for its quality by comparing it to other footprint sources, as any gaps in the chosen source will result in corresponding gaps in the final inventory.
4. **Synthesize data sources:** Synthesizing multiple data sources into a single inventory involves several steps: preprocessing data, attributing sources to the baseline geometry, selecting values for features with disagreement, filling gaps, and mapping to features required for simulation. This paper introduces the inventory synthesis framework shown in Figure 2 as a standardized approach for this process. The white boxes represent standardized steps in the workflow, whereas the colored boxes show an example of one way input data of various types could be used in the workflow.

Within the inventory synthesis framework, input data is categorized by source geometry, as different geometries require different methods for attributing data to the baseline geometry. The framework accommodates any combination of these input types. Furthermore, one input should be identified as the *bounding geometry*, which is a polygon geometry used to assist the process of attributing points to footprints. The bounding geometry polygons should exhaustive (cover the space completely), such as census block polygons or tax parcel polygons.

There are four main steps to preprocess data (Figure 2, Box A). First, all data sources must be trimmed to the appropriate study boundaries. While this may seem trivial, different data sources may have slightly different definitions of area boundaries, and it is important to be consistent to prevent issues in the later footprint attribution near the study boundaries. Second, individual data sources are cleaned by removing unreasonable building feature values. These can include negative population values (which can either be an error or a placeholder value), unreasonable year built data, or other features. Second, all sources must be mapped to a common ontology, which is a standardized vocabulary to classify building features in an inventory. For example, a multi-unit residential building might be categorized as RES3C in NSI, while tax parcel data may label it as "Condominium – townhouse style" or "Multiple residential building of 5 or more units." This step maps these source-specific values into a preferred vocabulary that allows different sources to be synthesized effectively in later steps. Finally, individual data sources can be enhanced as part of preprocessing, including linking non-geolocated sources to geolocated sources using other features, or dropping certain data if appropriate.

Beyond preprocessing, attributing sources to the baseline geometry is the next major step in the synthesis framework (Figure 2, Box B). This attribution process is completed in two steps. First, sources are tagged with

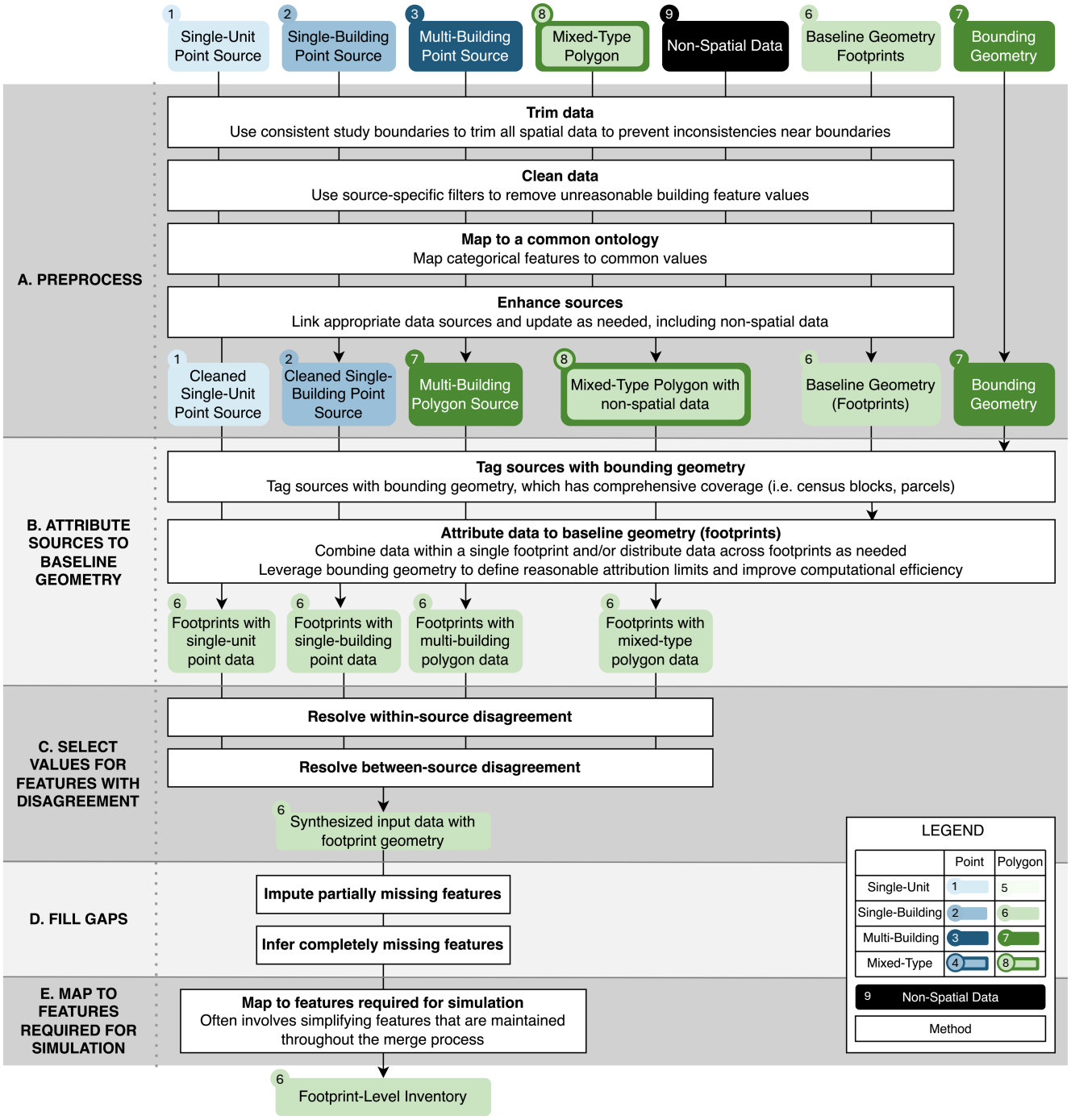

Figure 2: Inventory Synthesis Framework

a bounding geometry. For example, if census block polygons are being used as the bounding geometry, all points, footprints, and other sources would be tagged with the appropriate census block number. Second, data would be attribute to the baseline geometry, leveraging the bounding geometry to provide reasonable distance limits in the attribution process and improve efficiency. The result of this process should be that all sources are attributed to a single baseline geometry, which allows for information from each data source to be synthesized.

To produce a synthesized building inventory, each building feature for each baseline geometry (footprint) should be assigned a specific value. When multiple data sources are combined, it is common to find the same building feature in more than one data source. The corresponding feature values often disagree within each baseline geometry (footprint), and the next step of the framework is to select values for features with disagreement (Figure 2, Box

C). These disagreements can be characterized as either within- or between-source disagreements. *Within-source disagreement* occurs when multiple instances from the same data source are attributed to a footprint and they report conflicting building feature values. For example, two NSI points associated with one footprint may report different foundation types, which is not realistic in a building. Within-source disagreements can be resolved by comparing with the description of the same feature from an independent (additional) data source, or simply by selecting from the available feature values. *Between-source disagreement* describes when two different sources report inconsistent feature values for the same baseline geometry. This occurs when one or more sources contain incorrect information. For example, an NSI point and a tax parcel, both attributed to the same footprint, can report different numbers of stories. It is possible that the between-source disagreement is due to one or more data sources being disaggregated, meaning data may agree at an aggregated level (e.g., a census tract) but be inaccurate at the footprint level. These disagreements can be resolved by prioritizing one source over another.

In some infrequent cases, within- or between-source disagreements are not conflicts, but instead demonstrate the need for a more sophisticated building taxonomy. For example, if a single building is labeled as both a multi-family residential building and a commercial building, it is possible that it is a mixed-use building, and the vocabulary is not sufficiently granular for describing that complex occupancy type. We introduce the concepts of within- and between-source disagreement, but it is up to the user, based on context, to determine if these disagreements are due to inaccuracies or an insufficient taxonomy. In either case, it is important to acknowledge the uncertainty revealed by the disagreement in the data. Such uncertainty could be propagated using a Monte Carlo analysis that involves the generation of multiple possible building inventories and using these in downstream risk assessment to investigate how the outputs affected by the disagreement in the inventory data.

The next step of the synthesis framework is to fill gaps by approximating missing information for critical building features (Figure 2, Box D). We recognize two types of gaps: A feature is *partially missing* when it is available for a subset of the buildings, and it is *completely missing* when it is missing for all buildings. Partially missing information can be either random or systematic. There are several robust methods available to address randomly missing information through *imputation*, where missing values are estimated based on patterns in available data. However, data is often missing systematically in realistic inventories, and filling such gaps through imputation risks introducing bias. Such issues are best resolved through additional data collection or by treating the buildings with these gaps as if their features were completely missing. Completely missing features can be *inferred* by making reasonable assumptions about them using available information on other features. For example, the structural system of a building can be inferred if its height, occupancy type, and year of construction are available.

The final step of the framework is to map to the features required for simulation (Figure 2, Box E). In many cases, the information needed to run the risk simulation is slightly different than the building features maintained throughout the inventory development process. For example, while a risk assessment may only need 'design era,' it can be beneficial to track year built information throughout the inventory development process, and only map to design era right before the simulation. Similar mappings can include going from specific multi-family residential tags (e.g., 'RES3C' or 'RES3F') to the more general multi-family residential tag used in Hazus ('RES3'). By mapping features right at the end of the framework, more detailed data can be tracked throughout the process and potentially used for other types of analyses.

4. Methods and Illustrative Case Study

Based on the inventory synthesis framework in Figure 2, this section proposes specific methods for handling each step. Three different workflows based on the framework in Figure 2 are presented: 1) using only nationally available data, 2) using only locally obtained data, and 3) combining all available data sources, resulting in a "best estimate" inventory. The national workflow was designed to support broad application across the United States with minimal required changes. The methods presented in the local and best estimate workflows could also be applied in various locations, but using them in a new location requires a modest effort to develop scripts that handle the specific features and schema of local tax records and other location-specific datasets. We illustrate all proposed methods with a concrete example using data from Hayward, California, to provide a clearer explanation. Prior to synthesizing data, this section considers the key steps for developing a building inventory outlined in the previous section: selecting an appropriate resolution, selecting data sources, and selecting a baseline geometry.

Select appropriate resolution: First, it is important to determine the appropriate resolution based on the goals of the assessment. In the Hayward example, we assume the study targets a building-level seismic risk assessment using Hazus fragility curves. Thus, this study requires a building-level spatial resolution and typology resolution sufficient to inform the selection of a Hazus fragility curve. This requires building occupancy class, number of stories, year built, and structure type.

Select data sources: The Hayward example study draws from both national and local data sources. At the national level, it incorporates NSI data, as well as several HIFLD data sources. The latter are incorporated to complement NSI and provide higher-quality information for specific occupancy classes, such as robust locations of police, fire, mobile homes, and emergency response centers, as well as locations and populations of schools and college and university campuses. In this study, study, *mobile homes* will be used to refer to mobile (pre-1976) and manufactured (post-1976) housing, for consistency with the Hazus definition of the RES2 occupancy class. Additional national-level data sources include census block, tract, and place geometries, as well as housing unit and population counts from the 2020 Decennial Census. We also considered several building footprint data sources with national or global coverage, including OpenStreetMap, Microsoft Footprints, USA Structures, and Overture Footprints. At the local level, several data sources from the City of Hayward Open Data Portal were incorporated, including local property assessor parcel data (parcel geometries and extended metadata collected for property tax assessment purposes), building footprints, address point data (point geometries describing location, occupancy class, and other features), and zoning polygons, which regulate development within the city [65]. The California School Directory is used to supplement school features from HIFLD with building year-built data. These sources collectively contain all the required building features, except for structure type, which was not available from any source. This is a common limitation, as information on the structural system of buildings is typically not available in public datasets. Details on building feature terminology is shown in Appendix B. Data sources are summarized in Table 1, and more descriptive information is provided in Appendix D.

Table 1: Inventory data sources and the building features they provided for the Hayward Case Study.

Scope	Data Source	Date or Version Number	Geometry Type	Used in Case Study	Occupancy Class	# of Stories	Year Built	Population	Number of Units	Bldg Value	Bldg Type (Material)
Global	Overture Footprints	v.1.9.0 (2025)	Polygons (Building)		Sparse	Sparse					
	OpenStreetMaps	1/10/25	Polygons (Building)		Sparse	Sparse	Sparse				Sparse
	Microsoft Footprints	2019-2020	Polygons (Building)								
National	USA Structures	4/23/25	Polygons (Building)		*	Height					
	National Structures Inventory (NSI)	2022	Points (Building)	*	*	*	Median (Block)	*	Range	*	*
	HIFLD - Public and Private Schools	3/27/24	Points (Building)	*	EDU1			*			
	HIFLD - Colleges and Universities	12/16/22	Points (Campus)	*	EDU2			*			
	HIFLD - College and University Campuses	12/16/22	Polygons (Campus)	*	EDU2			*			
	HIFLD - Local Law Enforcement	12/10/23	Points (Building)	*	GOV2			*			
	HIFLD - Fire and EMS Stations	1/1/25	Points (Building)	*	GOV2						
	HIFLD - Emergency Operations Centers	12/10/23	Points (Building)	*	GOV2						
	HIFLD - Mobile Home Parks	8/14/23	Points (Park)	*	RES2				Total (Park)		
	Decennial Census Data	2010 Census	Census Block	*				Total (Block)	Total (Block)		
	Decennial Census Data	2020 Census	Census Block	*				Total (Block)	Total (Block)		
State	California School Directory	v.3.3.0.0 (2024)	Non-Geolocated	*	EDU1		Date Opened				
City	Hayward - Tax Parcel Data	6/26/18	Polygons (Parcels)	*							
	Hayward - Extended Parcel Data	2025	Non-Geolocated	*	*	*	*		*	*	*
	Hayward - Address Data	6/26/18	Points (Addresses)	*	*				Addresses per Footprint		
	Hayward - Footprint Data	6/26/18	Polygons (Building)	*		Height					
	Hayward - Zoning Data	6/26/18	Polygons (Zone)	*	Zone						

Note: An asterisk (*) indicates the feature is directly available. A description indicates one of the following cases: 1) Partial/Aggregate Data: Only partial (sparse) or aggregate (e.g., median, total) data is available, 2) The feature has a uniform value across the entire data source (e.g., "EDU2"), 3) Proxy Feature: A different feature can be used as a proxy for approximation (e.g., height).

Select baseline geometry: This step establishes building footprints as the baseline geometry. OpenStreetMap, Microsoft Footprints, USA Structures, Overture Footprints, and Hayward's local database were considered for use. Since no single data source represents the definitive "ground truth," all sources were visually compared to identify gaps and inconsistencies (as shown in Figure 3), using Google Maps satellite imagery to resolve discrepancies.

At the time of the study, OpenStreetMap and Microsoft Footprints, both accessed through the SimCenter BRAILS++ tool, had significant gaps in Hayward, especially in residential areas. This is particularly visible in Figure 3, Views A and B. USA Structures was the most comprehensive national data source but it was missing recently constructed buildings, visible in Figure 3, View C. Overture included newer construction and had more accurate geometries, but it had a few substantial gaps, particularly in areas with single-family residential buildings, shown in Figure 3, View B. The local footprint source was the most comprehensive and was selected for this case study. Since footprints are the baseline geometry adopted for the final inventory, selecting a reliable source is critical. Coverage patterns may differ elsewhere and evolve over time, so it is generally recommended to create comparison plots like Figure 3 to identify the best baseline geometry source. Future work may aim to combine

multiple footprint data sources, but due to the comprehensive coverage of the Hayward footprints, this was not included in the scope of this study.

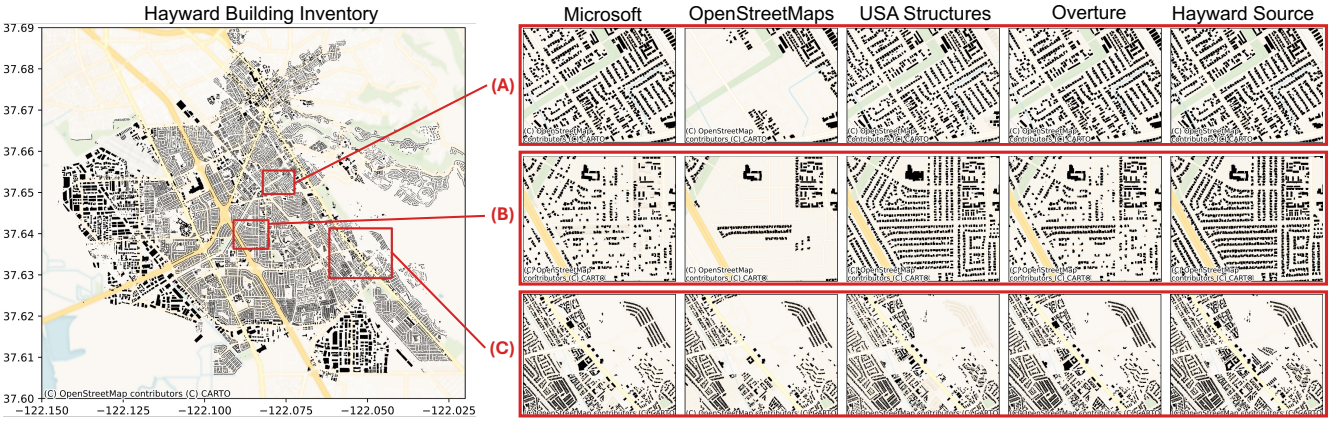

Figure 3: Coverage of five different footprint sources for different locations in Hayward.

Synthesize Data Sources: The final step is to synthesize selected data sources into a building inventory. The three workflows – national, local, and best estimate – are presented, and specific methods for attributing point and parcel data to building footprints are proposed alongside these workflows. Code corresponding to all methods is available on Github.

4.1. Nationally Available Data Source Synthesis

The following discussion demonstrates how the general inventory synthesis framework in Figure 2 is applied to create the workflow in Figure 4 that generates a footprint-level inventory using only nationally available data sources. Input data includes NSI, HIFLD, census block data, and Hayward building footprints (see Table 1 for details). The pink highlighted sections denote more detailed methods proposed in this study that are outlined explicitly in the text.

Preprocess Data (Figure 4, Box A)

The first step in preprocessing is trimming data to consistent study boundaries. Here, 2010 US Census Place designations are used to identify census blocks associated with Hayward and define the boundaries of the study area. Later analysis is based on census blocks, so the study area strictly follows the census block geometries, i.e., no partial blocks are included. The 2010 Census geometries are chosen to be consistent with the NSI data.

The next step is to clean data using source-specific filters. For the building footprints, this includes (1) removing footprints smaller than 450 square feet, a size threshold applied in the USA Structures dataset, as these typically correspond to sheds or other small buildings, and (2) removing overlapping and duplicate building footprints. In addition, negative population values are removed from all HIFLD data (-99 is often used as a placeholder). Finally, according to its documentation, NSI uses two identification fields: the fd_id ("a number that should be unique for all structures") and the *bid* ("building ID"). Many points in the NSI data share the same *bid* value. Based on spot checking several of these in Google Street View, it appears that points with the same *bid* belong to the same building. Thus, we grouped points by *bid* and condensed them by either summing their feature data (e.g., replacement cost) or storing it in lists (e.g., foundation type), as appropriate. Each *bid* group is treated as a single point in subsequent steps.

Mapping features to a common ontology is the process of standardizing data to a selected classification system to ensure consistency across the different data sources. In this study, a slightly extended version of the Hazus ontology is adopted. In the national workflow, there is only minimal mapping required because NSI data already uses the same ontology as Hazus to describe its features, including occupancy class, building type, and others. Thus, the only mapping done for the national workflow is to assign occupancy class values for the HIFLD data, and since each HIFLD dataset only describes one occupancy class, the mapping is straightforward. We extended the Hazus ontology to capture some of the more specific distinctions within the HIFLD dataset, such as subdividing the general 'GOV2' category to differentiate between fire stations ('GOV2-FIRE') and police stations ('GOV2-POLICE'). The complete list of extended occupancy classes is available in Appendix C.

The final preprocessing step is to enhance the data by linking appropriate sources and updating as needed. Whereas Box B of the national workflow aims to attribute data to building footprints, this step aims to just

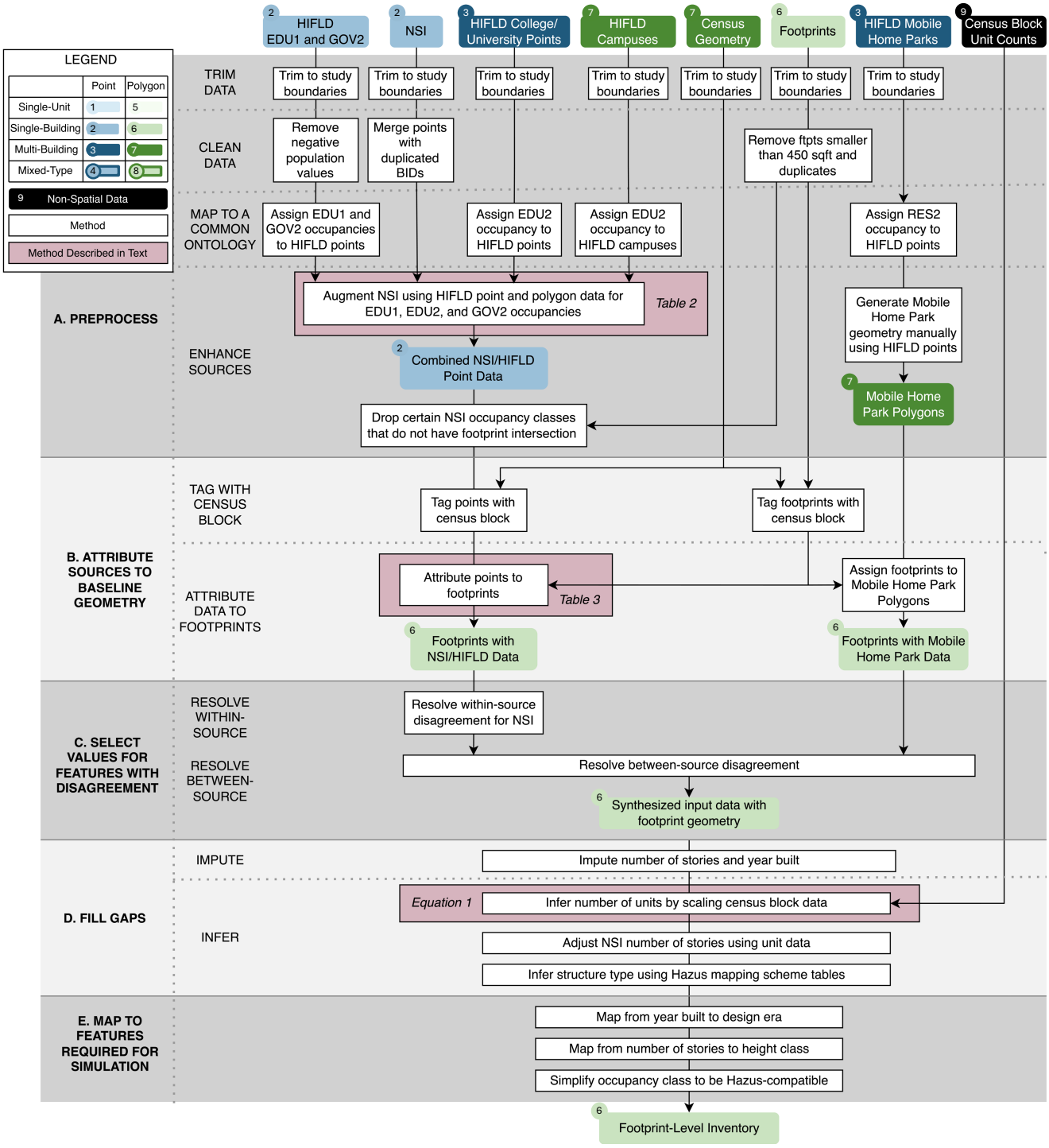

Figure 4: Inventory Synthesis Workflow using Nationally Available Data Sources

improve the quality of the data prior to synthesis. There are three specific methods used in the national workflow. First, the Hayward example study revealed that NSI does not reliably capture mobile homes, and this prompted the development of an alternative approach to identify these assets. The relevant HIFLD dataset provides a single point for each mobile home park, which is insufficient if the goal is to accurately assign footprints to this occupancy class. Thus, the dataset was enhanced by (1) importing the HIFLD Mobile Home Park points into Google My Maps, (2) manually drawing the boundary of each mobile home park as polygons based on satellite imagery, and (3) exporting these polygons and incorporating them into the national workflow.

Enhancing sources also involves augmenting and modifying NSI data using HIFLD point and polygon data for

the EDU1, EDU2, and GOV2 occupancies. This process, described in Table 2, was developed through evaluating the data in Hayward and spot checking with Google Street View. HIFLD appears to be a robust source of location information, but NSI has additional features available for these types of buildings. We observed in Hayward that EDU1, EDU2, and GOV2 points in the NSI data without an associated HIFLD point usually do not actually mark schools and/or emergency response buildings, and are thus dropped in this method. Furthermore, NSI GOV1 points are often clustered around HIFLD data points; it is unclear what these points represent and they may be an artifact of the inventory generation behind NSI. We recognize this by removing such these points.

Table 2: Method to augment NSI using HIFLD Point and Polygon Data for EDU1, EDU2, and GOV2 Occupancies

Step 1: Augment NSI with HIFLD GOV2 Point Datasets (Police, Fire, Emergency Operations)
Incorporate HIFLD GOV2 points directly into NSI as new GOV2 points. If an existing NSI GOV2 point is within 50m of a new HIFLD point, attribute its data to the new HIFLD point; otherwise drop remaining NSI GOV2 points not within 50m. Remove GOV1 points within a 10m of the new HIFLD GOV2 points.

Step 2: Augment NSI with HIFLD EDU1 Point Datasets (Public Schools, Private Schools)
Incorporate HIFLD EDU1 points directly into NSI as new EDU1 points. If an existing NSI EDU1 point is within 50m of a new HIFLD point, attribute its data to the new HIFLD point; otherwise drop remaining NSI EDU1 points not within 50m. NSI GOV1 points are often clustered around schools so, drop GOV1 points within a 50m of the new HIFLD EDU1 points. Assign HIFLD school population information as daytime population under the age of 65.

Step 3: Augment NSI with HIFLD EDU2 Datasets (Colleges and Universities)
Associate HIFLD College/University Points with HIFLD Campus Polygons by spatial intersection (overlap). Handle the following three scenarios:

- **HIFLD Campus with HIFLD Point + NSI GOV1 Points:** Convert all NSI GOV1 points within the campus polygon to EDU2 points. Proportionally scale up the campus population using the method below.
- **HIFLD Campus with HIFLD Point + no NSI GOV1 Points:** Incorporate HIFLD EDU2 points into the NSI data as new EDU2 points. Proportionally scale up the campus population using the method below.
- **HIFLD Point without Campus Polygon:** Incorporate HIFLD EDU2 points into the NSI data as new EDU2 points.

Method to Proportionally Scale Up NSI Campus Population:
In Hayward, the total NSI population within the campus polygons was much lower than the HIFLD-specified school population. Because HIFLD data is considered more reliable than NSI population counts, scale NSI campus populations proportionally to match HIFLD totals. To do so, set daytime population under age 65 to 99.5% of the HIFLD population and set daytime and nighttime population over 65 and nighttime population under 65 as 0.5% of the HIFLD population. Hazus also uses HIFLD to define demographic information about EDU2 points, and the allocation of population based on age and time of day is based on the Hazus 6.1 Inventory Manual.

Finally, it was observed in Hayward that there are certain occupancy classes that, if not exactly overlapping with a footprint, tended to be located outside of areas that contained footprints. More specifically, there were many instances of GOV1 points that were located far from any footprints out in marshland, and there were IND4 and IND5 points that consistently appeared along transit lines (see Figure 1c). Based on the quantity of these points, their presence or location may be an artifact of the way NSI is developed. Thus, GOV1, IND4, or IND5 points that are not explicitly located inside of a building footprint are dropped as part of the preprocessing stage.

Attribute Sources to Baseline Geometry (Figure 4, Box B)

To synthesize information across multiple sources, all sources must be attributed to a single baseline geometry (footprints). For the NSI and HIFLD datasets, a specific point-to-footprint attribution method was developed to automate these assignments. The goal of the point-to-footprint attribution method is to minimize data loss while assigning point data to building footprints. Standard methods such as spatial intersection joins often drop valuable data because the points do not explicitly overlap with a footprint. Dropping such points introduces bias and/or inaccuracies in the associated features and population in the inventory.

The proposed point-to-footprint attribution method is conducted in two steps; first, points and footprints are assigned to a broader bounding geometry, which is a geometry that is exhaustive (covers the space completely) and is used to improve the efficiency and scalability of the attribution process. Bounding geometries provide reasonable limitations for attributing data as well. In the case of the national workflow, census blocks are selected as the bounding geometry, and points and footprints are assigned to a census block based on intersection. Each census block is processed separately in the second step, when points are attributed to footprints in each census block. This process, described in Table 3, was developed through a detailed manual analysis of Hayward NSI data, supplemented by extensive review of Google satellite and street view imagery to resolve common issues and

challenges. Conversations with other researchers in the field suggest that the issues identified in this paper are common in NSI data and appear in other locations. Although designed based on NSI data, the approach explained below is later adopted for other point-to-footprint attributions.

Figure 5 illustrates the point-to-footprint attribution process for two example census blocks, following the steps outlined in Table 3. Figure 5a displays all points and footprints for the specified census block (Step 1). Figure 5b shows the points and footprints that are attributed in Steps 2 and 4. Figure 5c shows the remaining points and footprints beyond Step 4, with points dropped due to occupancy classed marked accordingly. This figure emphasizes that remaining points should not automatically be attributed to the remaining footprints. Instead, as shown in Figure 5d, footprints marked as "not full" may be more appropriate for absorbing these remaining points.

Figure 5: Point-to-footprint attribution method for two example census blocks

In the final step, remaining points are dropped from the footprint-level inventory. In the national workflow, points in this category were almost all in census blocks with no building footprints to begin with, such as census blocks that are along roads. An example of remaining points is shown in Figure 6. In Hayward, these unassociated points represent 247 buildings with a replacement cost of 403 million dollars and a nighttime population of 1,742. While the methodology aims to retain as many points as possible, it is unclear whether and where these points should be attributed, and they may be an artifact of the way the NSI data is generated and assigned to census blocks and regions. Based on review of the surrounding Hayward inventory, attributing these points to nearby footprints would have resulted in these footprints having an unrealistically high number of units, population, and value. Thus, these points are dropped from the inventory.

Select Values for Features with Disagreement (Figure 4, Box C)

If multiple sources describe the same building feature for a single footprint, a single value for that building feature must be selected. For each feature with disagreement, this means deciding which data point or data source to use to describe a feature, and in what order of preference. As part of this process, disagreements both within and between sources must be resolved. An example of the process is described here for occupancy class. No within-source disagreement is found in the HIFLD data, as no two HIFLD points are attributed to the same footprint. However, within-source disagreement can occur in the NSI data when multiple occupancy classes are attributed to the same footprint. In such cases, possible occupancy classes within a single building footprint are grouped into several classes and prioritized using the following order: 1) educational and emergency response buildings, 2) large residential (RES3C-RES3F), 3) non-standard residential (RES4-RES6), 4) small residential (RES1, RES3,

Table 3: Point-to-footprint attribution method (Figure 4, Boxes B4 and B5)

Step 0: Assign features that should be summed versus stored as lists
Throughout this method, many cases involve attributing multiple points to a single footprint. To facilitate this, assign all building features into one of two categories: 1) features that should be summed across points attributed to the same footprint (i.e., dollar value, population) and 2) features that cannot or should not be summed (i.e., foundation type, number of stories). Throughout the attribution process, features that cannot be summed will be stored as a single value if all points agree on the feature, or stored as a list of possible values if points do not agree on the feature.

For each bounding geometry (i.e., census block, parcel geometry):

Step 1: Select relevant points and footprints
Select points and footprints associated with the target bounding geometry.

Optional: Include additional footprints from adjacent bounding geometries in the attribution process. Footprints from adjacent blocks can help handle cases where nearby points and footprints fall in different bounding geometries or if a footprint spans across a boundary.

Step 2: Attribute trivial cases with one point intersecting with one footprint
Attribute points to footprints in cases where a single point intersects a single footprint.

Step 3: Set gross area limits by occupancy class
This step introduces a threshold limit on gross area for each occupancy class to avoid attributing too many points to a given footprint. Determine the threshold by computing the mean and standard deviation of the square footage, calculated as the product of the footprint area and number of stories (estimated if unavailable) for all point/footprint pairs defined in Step 2. For each occupancy, define the threshold as two standard deviations above the mean. Footprints that have an associated gross area below the threshold for the occupancy class are marked as "not full," indicating potential to accommodate additional points.

Optional: Use a gross area limit based on occupancy class. A gross area limit can be helpful if working with a dataset that is largely single-building point data, where footprints should be de-prioritized if they already contain points. A limit is not appropriate if a dataset contains mostly single-unit point data, where a single footprint should absorb many points, by design.

Step 4: Attribute cases with multiple points intersecting with one footprint
Attribute points where multiple points intersect a single footprint. Also attribute points to footprints that do not intersect, but fall within a courtyard of a footprint by extracting the interior rings (holes) in each footprint and converting them to polygons representing enclosed courtyard spaces, allowing points located in these voids to be properly attributed to the surrounding footprint. When multiple points are attributed to one footprint, combine data across points by either summing or storing lists of values, per Step 0. In addition, flag unusual occupancy class combinations for manual review using Google Street View (e.g., single-family residential and metals/minerals processing in the same footprint).

Optional: Update residential occupancy classes to reflect combination of multiple points (e.g., two single-family (RES1) points in the same footprint get converted to a duplex (RES3A) point). If unit count is not explicitly provided for point data, estimate the number of units based on the average of the unit range defined for each multifamily (RES3) type in the Hazus Inventory Technical Manual [55]. These updates should be used when working with single-building point data (where points expressly represent different units) and should not be used if working with single-address point data, where each point represents one unit, but occupancy is often tagged for the entire building (i.e., one unit in a 8-unit building would still be tagged as RES3C).

Step 5: Attribute points nearby to footprints, based on a threshold distance
Attribute remaining unassigned points based on proximity, using a threshold distance below which a point is 'nearby' a footprint. In this study, 10 m was used as the threshold. If the gross area limits set in Step 3 *are not* being used, attribute points to their closest footprint (within the 10 m limit), whether or not there are existing points already in the footprint from Steps 2 or 4). If the gross area limits set in Step 3 *are* being used, attribute points to footprints in the following order, thus prioritizing the assignment of points to empty and "not full" footprints.

1. Empty footprints: If a point is within 10 m of an empty footprint, attribute it to the closest empty footprint.
2. "Not full" footprints: If a point is within 10 m of a footprint marked as "not full," assign it to the nearest "not full" footprint, combining features using the methods from Step 4.
3. Any footprint: If neither condition applies, assign the point to the nearest footprint, combining features using the methods from Step 4.

Optional: Include footprints from adjacent bounding geometries in the attribution process. Similar logic to Step 1 can be applied here.

Step 6: Attribute points farther from footprints, based on a threshold distance
Using the same approach as Step 6, assign remaining points based on proximity, using a farther distance threshold (100 m in this study).

Optional: Include footprints from adjacent bounding geometries in the attribution process. In this step, it is typically better to limit the considered footprints to only the bounding geometry of interest (not adjacent).

Step 8: Drop remaining points
Remove any remaining unassigned points from the inventory.

RES3A-RES3B), and 5) other occupancies not included in these categories. Information is prioritized in this order

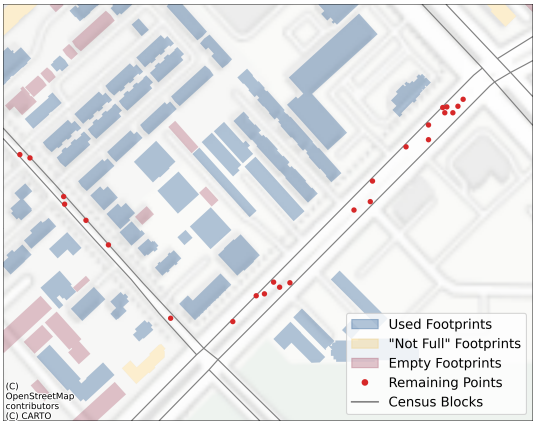

Figure 6: NSI points located along road-only census blocks, dropped at the end of the workflow

to ensure that buildings that may have higher impact (a 50+ unit structure) are not overwritten by another possible option in the footprint. If a residential occupancy is paired with a non-residential occupancy, a "M" is appended to the residential occupancy class to indicate a mixed-use footprint (e.g., 'RES3CM'), further extending the Hazus ontology. Finally, between-source disagreement occurs only when both HIFLD and NSI data are assigned to the same footprint. Since the HIFLD data appears to be more reliable through spot checking points in Hayward, it is prioritized over NSI when occupancy class disagrees. Similar procedures are applied for other building features (besides occupancy class) as well.

Fill Gaps (Figure 4, Box D)

Due to its comprehensiveness, there are no features that are partially missing at random in the NSI data. However, HIFLD data is systematically missing year built, number of stories, and building type; that information is simply not included the HIFLD datasets. SimCenter's imputation methods are used to address these partially missing features [18]. In general, imputation is best suited to impute data that is missing at random and thus can be filled in by patterns in surrounding data. The HIFLD data does not represent random gaps, but rather specific community assets, which makes imputation less appropriate. However, as neighborhoods and cities develop, key community building services – such as schools and fire stations – often follow similar patterns of development. Thus, imputing the year built for HIFLD points based on surrounding buildings seems appropriate. Additionally, since most construction in Hayward is low-rise, imputing the number of stories likely has minimal impact on Hazus fragility curve selection.

In addition, the NSI public fields do not provide exact unit counts for residential buildings, only ranges. For example, a RES3D building is defined as having 10-19 units. This study explored several methods to assign more specific unit counts. During this process, it became clear that the NSI data overestimates the number of RES3F (50+ unit) structures, a pattern that was observed in Hayward and confirmed to be observed in other regions by the Hazus 6.1 developers. To correct for this overestimation and get an explicit unit count, the following scaling factor (informed by the 2020 Census) is multiplied by the population of each footprint in the census block, and the result is rounded to estimate the number of units per footprint. RES1, RES2, and RES3A buildings are assumed to have 1, 1, and 2 units, respectively, and are not adjusted. A lower bound of 2 units is used when scaling multi-unit residential buildings. This process enables the workflow to provide explicit unit counts and correct the overestimation of RES3F points in NSI using census data.

$$scaling\ factor = \frac{U_{total} - U_{RES1,RES2,RES3A}}{P_{RES3B-3F}} \tag{1}$$

U_{total} : Total number of units in census block based on 2020 Decennial Census

$U_{RES1,RES2,RES3A}$: Total number of RES1, RES2,and RES3A units in the census block from NSI

$P_{RES3B-3F}$: Total population in RES3B to RES3F structures in the census block from NSI

Finally, neither NSI nor HIFLD report the structural system, which is required to select the appropriate Hazus fragility curve. Thus, this feature is completely missing and must be inferred. The structure type is inferred using the Hazus building stock mapping scheme tables [66], which provide distribution percentages of floor area for specific

structure types, given year built, number of stories, and occupancy class [9]. Three modifications were made to the Hazus structure type assignment. First, structure types for 2–4-unit residential buildings were sampled as if they were single-family homes, since these buildings are typically constructed with methods and materials more similar to single-family homes than to larger multi-unit structures. For example, it's unlikely for a 2-unit duplex to be a steel structure, which could otherwise occur if all multi-family (RES3) buildings are sampled the same, regardless of the number of units. Second, the MH structure type in Hazus was only assigned for mobile homes (RES2), rather than being a structure type option for multiple multi-family (RES3) occupancies. Finally, no URM structure types were assigned in Hayward because of the 1986 unreinforced masonry retrofit mandate, which requires local governments in high seismic zones to identify and establish loss reduction programs for URM buildings [67]. In addition, although building type (material) is available in the NSI, it was found to severely overestimate the presence of masonry, and the manufactured housing 'MH' building material did not correspond well with actual mobile homes. Thus, structure type assignment was done without regard to the NSI building type. For example, even if a NSI residential building is listed as having a masonry building type, possible structural systems are not limited to only masonry structural systems. Overall, inferring structure type introduces a significant amount of uncertainty due to the variability in structural systems across occupancy classes and the absence of standard methods to collect and report this information. Future work is needed on developing more granular, location-specific methods for structure type assignment.

Map to Features Required for Simulation (Figure 4, Box E)

The inventory in this study hypothetically targets a regional simulation using Hazus fragility functions. Based on the available features in the synthesized inventory, three mappings are needed to obtain features required for simulation. The numeric year built and number of stories values are mapped to the corresponding design era and height class categories. Additionally, the extended occupancy class, which allows for more detailed descriptions than Hazus (e.g., 'GOV2-POLICE', 'RES3F') are mapped to a more simplified ontology (e.g., 'GOV2', 'RES3'). This step completes the inventory synthesis workflow using national data and results in a building inventory that can be used for downstream regional risk assessment studies.

4.2. Locally Obtained Data Source Synthesis

The following discussion demonstrates how the general framework in Figure 2 is used to create a footprint-level inventory using local data sources. The local workflow for Hayward is shown in Figure 7. Input local data sources include tax parcel data (parcel property extent polygons and extended tax parcel data), address point data (geolocated points with information on each address), zoning data (polygons describing land use regulations and permitted development), and building footprints. In general, the development of a building inventory using only local data is not recommended since local tax data may be missing features that can be readily identified by integrating local and national databases. In this study, the inventory based only on local data was created as a separate point of comparison with the national workflow to identify key differences and potential biases in the data.

Preprocess Data (Figure 7, Box A)

The same preprocessing steps used previously – trim to the study area, clean data, map to a common ontology, and enhance sources – are used in the local workflow. More specifically, the same footprint cleaning process is used in the local workflow as was used in the national workflow to remove small and duplicated footprints. In addition, local data is cleaned by removing unreasonable values, which can arise due to typos or other errors in the data.

Mapping local data to a common ontology is more complex than for national data due to greater variability in parcel, address, and zoning descriptions. For example, the Hayward tax parcels contain 127 different text descriptions of the building use (occupancy class). To convert local data to a Hazus-compatible inventory, each description must be mapped to an NSI occupancy class. While some mappings are straightforward (e.g., 'Single-Family Dwelling' to RES1), others are more challenging (e.g., 'Miscellaneous improved commercial'). Hazus 6.1 documentation informs the mappings, but in the Hayward study, three key adjustments are made. First, mixed-use designations, which are not included in NSI or Hazus, are added to accommodate descriptions such as 'Multiple-Res building of 5 or more units + commercial units.' Second, some parcels and address points have descriptions such as 'Golf Course,' 'Pump,' or 'Traffic Signal', which are labeled in this workflow with the tag 'NOTBLDG'. It is possible that misattributing these types of parcels as structures may be the reason there are NSI points along roads and transit lines. Third, the occupancy class list is expanded to accommodate descriptions of either vacant or planned parcels, such as 'Vacant industrial land,' and 'Single-Family Residential – Planned Development Tract with Common Area'. These were marked with an additional 'VAC' following the otherwise appropriate occupancy class. All extensions to the NSI Occupancy Class ontology are outlined in Appendix C.

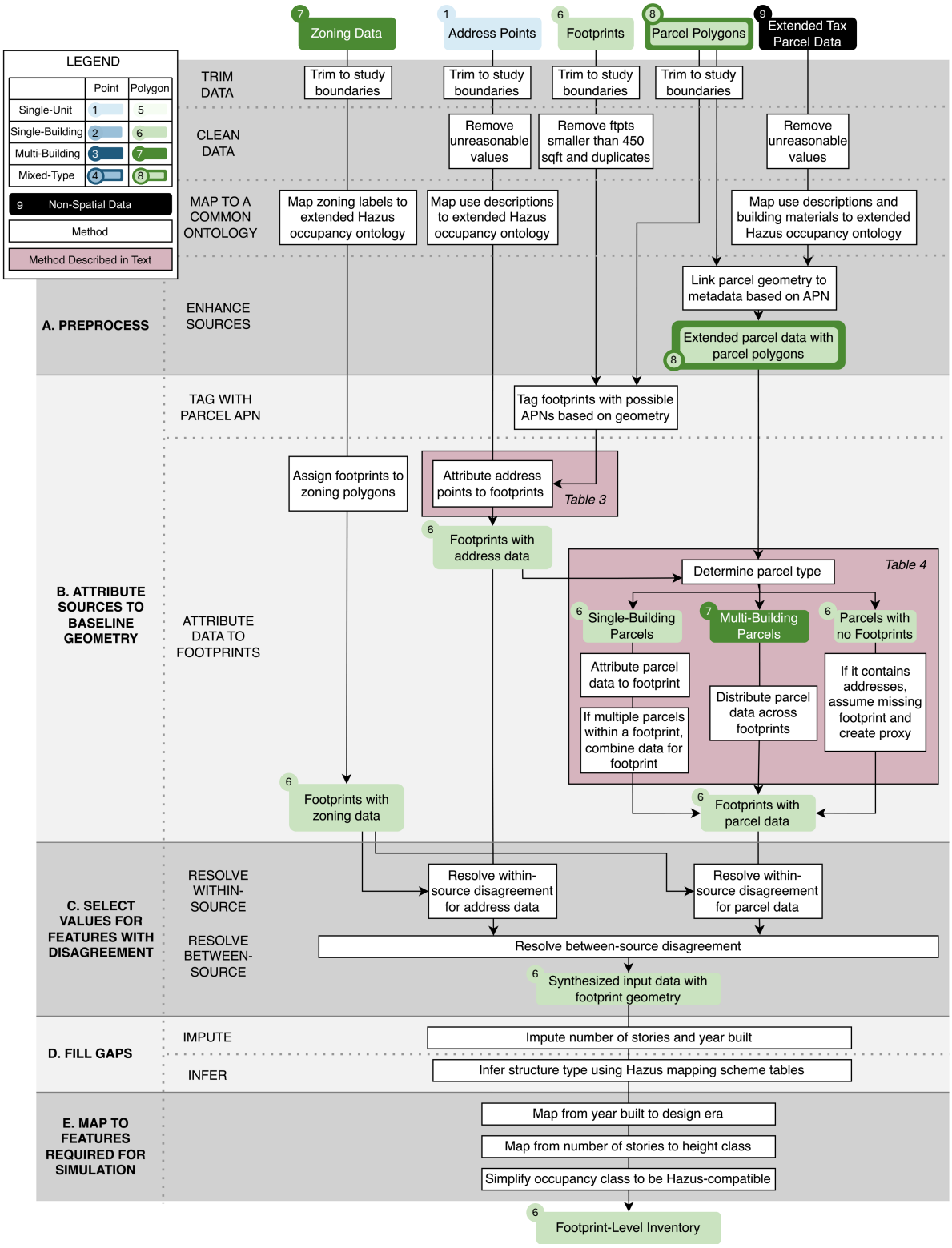

Figure 7: Inventory Synthesis Workflow using Local Data Sources.

Finally, parcel sources were enhanced by linking the parcel geometries to the extended tax parcel data from Hayward, which is non-geolocated. Geometries are linked to metadata using the Assessor Parcel Number (APN), which is listed in both datasets.

Attribute Sources to Baseline Geometry (Figure 7, Box B)

Similar to the national workflow, all sources are attributed to the baseline geometry (building footprints) using a two step process; first, data is tagged with an appropriate bounding geometry, and second, data is attributed to footprints. In the local workflow, the selected bounding geometry is the parcel polygons. Thus, all sources must be tagged with the bounding geometry the corresponding parcel number (APN). The address point source from Hayward already lists the appropriate APN for the point, so no additional tagging is required. Footprints are assigned a parcel APN if any part of the footprint overlaps with the specified parcel. For cases in which 95% of a footprint's area is within one parcel, other parcel associations for that footprint are dropped. Once all sources have been tagged with the appropriate bounding geometry, data is attributed to footprints efficiently by processing each bounding geometry separately. Address points are attributed to footprints using the same point-to-footprint attribution process described in the national workflow (Table 3). In addition, a parcel-to-footprint attribution process was also developed for the local workflow and is outlined in Table 4, with examples of each case shown in Figure 8.

Table 4: Method to attribute tax parcels to building footprints

Case	Assignment Criteria	Attribution Method	Example
1	APN corresponds to one footprint containing address data	Attribute all parcel data to the footprint containing address data. For cases where there are multiple parcels associated with a single footprint (Figure 8c), store parcel data either by summing (dollar value, etc.) across parcels associated with the same footprint, or by storing single values or lists of values for features that should not be summed (i.e. number of stories, building type, etc.)	Single-family home, or condominium unit within a larger building
2	APN corresponds to multiple footprints containing address data	Divide parcel data across footprints by either assigning the feature value to each footprint in the parcel (i.e. occupancy class), or by dividing data across all footprints in the parcel (i.e. dollar value). When data is divided, allocate it proportionally to the total area of each building	Apartment complex with multiple buildings under a single owner, or mobile/manufactured home park under a single owner
3	APN corresponds to one or multiple footprints, but no address points	If there is one footprint, attribute parcel data to the footprint. If there are multiple footprints, attribute the parcel data to each footprint using similar methods to Parcel Case 2	Industrial land with miscellaneous improvements (no corresponding address)
4	APN corresponds to one or multiple address points, but no footprints	If there is one address point, attribute parcel data to the point. If there are multiple address points, attribute the parcel data to each point using similar methods to Parcel Case 2. Then, group all address points without footprints into "likely footprints" using the following clustering approach: 1) For each address point within a parcel with no footprints, identify its "close address points" as those within a 7-meter radius, 2) Recursively link address points that share any "close address points" to form clusters, 3) Treat each resulting cluster as a proxy for a single building footprint, and combine address points accordingly	New construction, where footprints are not available yet, but the buildings are present in the tax records
5	APN has no footprints and no address points	Drop parcel data	Parking lot or water meter

Figure 8 shows an example the cases described in Table 4. Case 1, where a parcel corresponds to one footprint with address data, can include single-family homes (8a), single-family homes with sheds or garages that do not have addresses (8b), or a single unit within a larger building (8c). In these cases, there can be multiple parcels per footprint, but they still only have one footprint per parcel, and thus are Case 1. Case 2, where a parcel corresponds to multiple footprints with address data, can vary widely between just a few footprints (8d) to many footprints with many addresses, such as a mobile home park under a single owner (8e). Figure 8f shows an example of Case 3, where a parcel corresponds to one or more footprints, but does not correspond any address points. Finally, Figure 8g shows several examples of Case 4, where there is a parcel and one or more address points, but no footprint available. Based on the characteristics of these parcels, many of these cases represent new construction, which may not have footprints available yet.

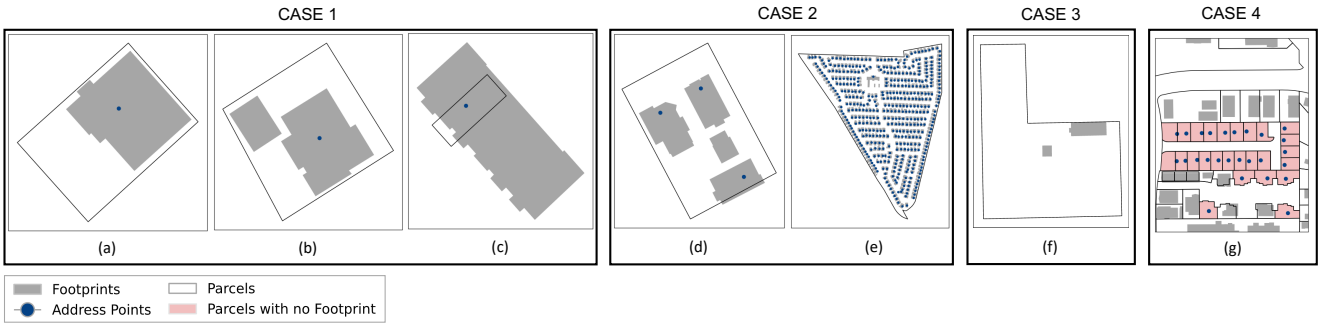

Figure 8: Examples of parcel, address point, and footprint geometries that correspond to each case described in Table 4.

Select Values for Features with Disagreement (Figure 7, Box C)

Prioritizing sources for features with disagreement is more complex with local data than national data for two reasons. First, local data, being manually generated, is prone to typographical and other errors. Second, while only one dataset had comprehensive coverage nationally (NSI), two sources (tax parcels and address points) have comprehensive coverage locally, leading to higher levels of between-source disagreement.

Possible typographic errors are addressed by screening for unreasonable values, and within-source disagreement is generally resolved by using the mode of each feature. For related features, such as *year built* and *effective year built*, values are resolved together to preserve relationships. For occupancy class, within-source disagreements in address point and tax parcel data are resolved using zoning data as an additional reference. In cases of conflict, the occupancy class that aligns with zoning specifications is prioritized. This use of zoning data provides a more robust way to resolve within-source disagreement, when otherwise there is limited information by which to do so. Between-source disagreements are also resolved. Since tax parcels can represent multiple buildings, while address points typically represent just one, address points are considered more footprint-specific and are prioritized when available. However, because address points do not include all features, parcel data is used when necessary. Footprints are removed if both address and parcel sources indicate that the parcel represents a vacant or planned lot or is not a building.

Fill Gaps (Figure 7, Box D)

The process to impute and infer data to fill gaps closely follows the approach used in the national workflow. SimCenter methods are used to impute gaps in year built and the number of stories, which are missing in the local data at rates of about 4% and 8%, respectively [18]. The main difference with the national workflow is that partially missing features in the local data seem to be missing at random, making this a better candidate for imputation. As in the national workflow, structure type is inferred using Hazus building stock mapping scheme tables [66]; however, in this case, the building type (material) from the tax data is used to inform which structure types are sampled for each building. The number of address points attributed to each footprint provides a robust estimate of the number of units per residential building; thus, the number of units is not inferred using census data for the local workflow.

Map to Features Required for Simulation (Figure 7, Box E)

The mapping for the local workflow closely follows that of the national workflow.

4.3. Best Estimate Inventory: Synthesis using National and Local Data Sources

The final workflow uses both national and local data to produce the best estimate inventory of Hayward. It incorporates all previously described inputs, along with the California School Directory, a non-geolocated dataset used to supplement HIFLD.

In comparing the results of the national and local workflows, it became apparent that a significant number of residential footprints only contained NSI point(s) and did not have corresponding address points or tax parcels. Further analysis revealed that the footprints that contained only residential NSI data (without local data) do not correspond to actual residential buildings but instead represent covered parking awnings or other features, as shown in Figure 9. There were 849 such footprints across Hayward, and the NSI points attributed to these footprints represented 416 million dollars and 10,370 people erroneously associated with these footprints. One potential solution is to drop the incorrectly attributed residential footprints, but doing so would also drop associated NSI points, significantly reducing the value and population captured in the inventory. However, leaving them would

cause people and value to be misplaced and structures to be mislabeled, biasing residential data. The best estimate workflow resolves this by updating the baseline geometry by removing these footprints from the data.

Figure 9: (a) Residential footprints containing both NSI and local data (blue) or containing only NSI data (red). Based on the Goggle Maps satellite image shown in (b), footprints containing only NSI data seem to be covered parking structures, not actual residential buildings.

Based on this update, the best estimate workflow can be visualized using Figures 4 and 7. First, Steps A and B of both the national and local workflows are run to preprocess data and attribute all sources to the baseline geometry (footprints). Then, results are compared to identify and drop footprints that contain NSI residential data but *do not* compare local residential data. A similar adjustment is made to the baseline geometry to incorporate proxy geometries (5-meter circles) for missing footprints like those in Figure 8g. The updated baseline geometry source is then used as an input into the national workflow to re-attribute national data to footprints (Steps A and B). This allows for the value and population to be absorbed into nearby multi-unit buildings, without dropping data or incorrectly attributing data to unrealistic structures. While this may not affect aggregated loss metrics, these nuances become more important if a study focuses explicitly on multifamily housing and/or aims to provide specific results to stakeholders.

The other change made in the best estimate workflow is the incorporation of the non-geolocated California School Directory, which is linked with HIFLD Public and Private schools based on school name as part of the *Enhance Sources* part of preprocessing data. The opening date in the directory was used to estimate year built, which is otherwise unavailable in the HIFLD data. It appears that 1980 is a default year for all schools that opened before 1980, so those entries are not adopted.

As a result of Steps A and B, with the above changes, both national and local sources have been attributed to updated building footprints. The next step in the workflow is Step C, which is selecting values for features with disagreement. Here, the integration of a larger number of data sources increases the complexity of this step. Within-source disagreements are handled as in the previous national and local workflows; however, increased levels of between-source disagreement must be resolved. In general, values are selected in the case of disagreement following a hierarchy. HIFLD information is prioritized first due to its specificity and accuracy. Address points are prioritized next because they are specific to individual footprints, followed by tax parcels, which have more features available. NSI is used when other sources are not available. Both local sources are prioritized over NSI because they are manually developed rather than generated through national-level assumptions, making them more reliable at the footprint level. Figure 10 illustrates occupancy class prioritization. Occupancy class is the only feature with no missing data because HIFLD, NSI, parcel, and address data all specify occupancy class, and footprints are dropped if they don't contain any of those four sources. HIFLD data is limited to select structures, so the majority of occupancy class assignments come from address points. Figure 10a shows a school assigned by HIFLD, residential buildings assigned by address points, and smaller buildings around the school, which don't have local data, assigned by NSI. Though address point data is available for most residential structures, Figure 10b highlights greater variability in industrial and commercial areas, where occupancy class is assigned from mixed sources based on availability.

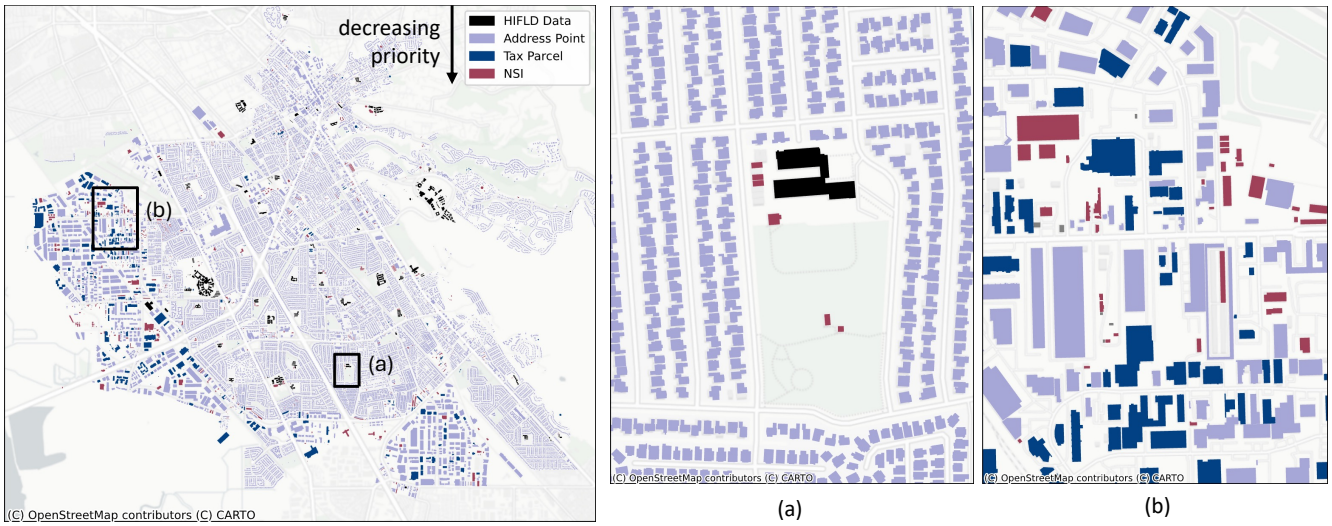

Figure 10: Example of Source Prioritization for Occupancy Class.

Imputing and inferring gaps in the data follow a similar process to the national and local workflows. If the building type is available from local data, it is used to inform which structure types should be sampled. If not, all structure types for the given occupancy class are sampled, following the approach used in the national workflow. The number of units is not inferred using census block data and no adjustment of NSI number of stories is needed (as shown in Figure 4) because more robust estimates are available from local data sources. The same three steps for mapping to features required for simulation are used in the best estimate workflow.

5. Implications and Illustrative Results

Through comparisons of summary statistics and individual building features including occupancy class, number of housing units, and building type, this section demonstrates that the choice of data sources and synthesis methods can significantly influence the resulting inventory. Five inventories are featured in this section. Three of them are produced using the workflows described in the previous section and are referred to as the *Synthesized National*, *Synthesized Local*, and *Best Estimate* inventories. Since NSI is used as the only data source in several studies, the raw NSI point data, referred to here as the *NSI Point* inventory, is also evaluated. A fifth inventory created by spatially joining the raw NSI point data and building footprints using intersection is also considered. This *NSI Spatial Join* inventory enables comparisons at the footprint level that are not possible with the *NSI Point* inventory. This approach also represents the default inventory generation workflow in SimCenter's BRAILS++ tool, v4.1.0 [18].

When aggregated at the city level, most summary statistics are similar across the five inventories. All inventories have a similar total number of buildings reported, other than the *NSI Point* inventory, which reports roughly 7,000 additional buildings. Similarly, all inventories report a similar total nighttime population, other than the *NSI Spatial Join* inventory, which reports roughly 20,000 fewer people. These differences are illustrated in Figure 11. The difference between the *NSI Point* and *NSI Spatial Join* inventories is because all NSI points that do not intersect a building footprint are dropped during the spatial join process. Thus, the *NSI Point* inventory may overestimate the number of buildings, but by dropping many data points and the corresponding population, the *NSI Spatial Join* inventory then underestimates the population. The synthesized inventories provide a more stable estimate because data is attributed to footprints, which stabilizes the building count, but minimal data is dropped in this process, thus stabilizing the population count.

It is important to note that the number of buildings and population in the *NSI Spatial Join* inventory are largely dependent on the selection of the footprint source. Using the robust Hayward footprint data source, 7,125 points are dropped when the NSI data is spatially joined with the footprints. However, even more data is lost if a less comprehensive footprint source is used. For instance, there are 9,742 fewer buildings in the inventory when NSI point data is spatially joined with OpenStreetMap footprints—about 24% of the total inventory. These findings further emphasize the need for careful consideration when selecting a baseline geometry source and attributing other geometries to it.

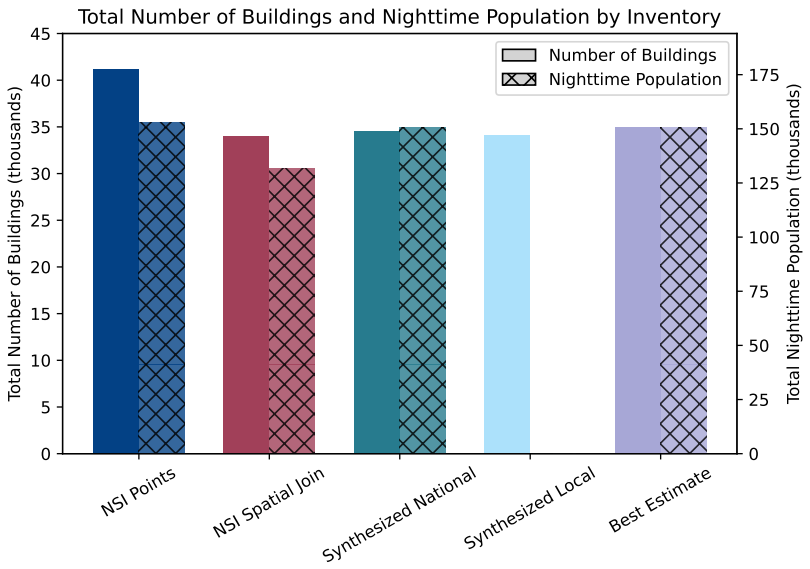

Figure 11: Total number of buildings and total nighttime population by inventory. The *Synthesized Local* inventory does not have a population estimate because there was no local source with population data available.

General patterns in occupancy class are similar across inventories, but there are a few important differences shown in Figure 12a. For example, the *NSI Point* inventory has a much higher number of residential buildings relative to other inventories; this is likely due to instances of multi-unit buildings that NSI characterizes as many RES1 points instead, artificially raising the residential building count. In addition, NSI contains more industrial (IND), commercial (COM), and government (GOV) buildings compared to local data sources. Many of these additional points are placed along transit lines and in locations with no buildings, suggesting they may correspond to nonstructural tax parcels (e.g., labeled as "Pump" or "Traffic Signal") mistakenly classified as buildings. Furthermore, both the *NSI Point* and *NSI Spatial Join* inventories underrepresent educational buildings, including schools (EDU1) and universities (EDU2). This is particularly dramatic for the *Spatial Join* inventory, which is missing a large percentage of schools because many NSI EDU1 points do not overlap with footprints, and thus were dropped in the spatial join process. Some NSI GOV1 points appear to be public housing or educational facilities, potentially contributing to the inflated count of GOV buildings and the underrepresentation of EDU buildings in both the *NSI Point* and *NSI Spatial Join* inventories. Local data sources also lack comprehensive coverage of educational and governmental buildings; some schools are incorrectly mapped to district office locations, while others are entirely missing from tax records, likely due to their tax-exempt status. The HIFLD datasets provide the most complete representation of governmental and educational buildings and are included in both the *Synthesized National* and *Best Estimate* inventories.

Moving beyond aggregated statistics, the number of footprint-level disagreements increases as more detailed building features are examined. For example, Figure 12b illustrates that disagreements are less common when using broad occupancy class categories (e.g., RES, COM), whereas Figure 12c shows more widespread disagreement when comparing specific occupancy classes (e.g., RES1, RES3A, COM2). Figure 12b and 12c also demonstrate that disagreements between different data sources do not occur at random but tend to be clustered throughout the city. Such clusters of errors in building occupancy type may cause bias in the results of regional risk studies.

Further insights emerge when focusing specifically on residential buildings, including single-family housing (RES1), mobile homes (RES2), and multi-family housing (RES3A-RES3F). Figure 13a shows that patterns in occupancy class counts are similar across inventories when aggregated at the city level, where most residential buildings in Hayward are single-family (RES1), with much fewer multifamily buildings (RES3A-RES3F). However, there are significant disagreements in the number of single-family homes (RES1), mobile homes (RES2), and residential buildings with 50+ units (RES3F). Compared to other inventories, the *NSI Point* inventory greatly overestimates the number of single-family homes, likely due to mistakenly characterizing multi-unit buildings as many individual single-family (RES1) buildings. In addition the *NSI Point* inventory overestimates the number of residential buildings with 50+ units (RES3F) and underestimates the amount of mobile homes (RES2) relative to local data.

Figures 13b and 13c illustrate that disagreements in footprint-level occupancy class values for residential buildings tend to cluster geospatially. Figure 13b compares the *NSI Spatial Join* against the *Synthesized Local* inventory,

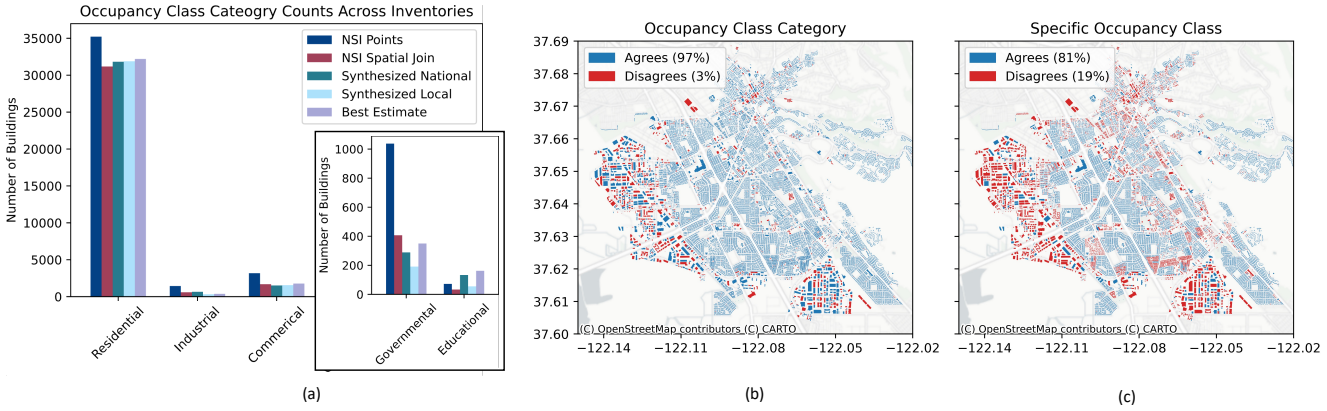

Figure 12: (a) Overall building counts by occupancy class category in each inventory. Footprint-level agreement of (b) general occupancy class category and (c) specific occupancy class between *NSI Spatial Join* and *Synthesized Local* inventories.

while Figure 13c compares the *Synthesized National* against the *Synthesized Local* inventory. Although both comparisons show disagreements, the *Synthesized National* inventory has better agreement with the local data than the *NSI Spatial Join* inventory. This demonstrates that careful synthesis of additional datasets provides more accurate approximations of local data. The most obvious improvement is the incorporation of mobile home park geometries, which cluster in the southern part of the city.

Figures 13d and 13e show confusion matrices for footprints where the inventories disagree on the residential occupancy class. The confusion matrices illustrate that synthesizing multiple data sources can reduce both the *quantity* and *severity* of disagreements. For example, many duplex buildings (RES3A) in the *NSI Spatial Join* inventory are labeled as mobile homes (RES2) in the *Synthesized Local* inventory (Figure 13d), which can lead to the selection of a significantly different vulnerability model, resulting in a biased seismic risk estimate by the *NSI Spatial Join* inventory. In contrast, the disagreements in Figure 13e between the *Synthesized National* and *Synthesized Local* inventories mostly involve duplex (RES3A) vs single-family (RES1) classifications. Disagreements between RES1 and RES3A occupancies are less consequential in this study because, as mentioned in the previous section, we use identical assumptions to infer structure type for these occupancy classes, recognizing that these buildings often have similar structural systems. This assumption relaxes the accuracy needed in classification between RES1, RES3A, and RES3B occupancies.

The *Synthesized National* residential inventory is a significantly closer approximation to local data than the *NSI Spatial Join* inventory for two main reasons. The first is related to the incorporation of mobile home park geometries. Improvements in mobile home (RES2) counts are demonstrated with three example census blocks in Table 5. The number of mobile homes from tax parcel data, HIFLD data, and manually generated polygons are in good agreement. In contrast, the *NSI Point* inventory and the inventory in Hazus 6.1 [55] are also in agreement and underestimate RES2 counts by a factor of 2 to 3. These disagreements point to a wider problem of systematic undercounting of mobile homes, which can result in underestimating the seismic risk to housing that is highly vulnerable to earthquakes and often occupied by more socio-economically vulnerable households.

Table 5: Number of Manufactured Homes (RES2) by Source for Three Example Census Blocks

Data Source	**Census Block 06001438204**	**Census Block 06001438201**	**Census Block 06001437200**
Hayward Tax Data	801	457	340
HIFLD	800	462	367
Synthesized National Inventory (Manually Generated Polygons)	806	461	379
NSI Points	244	123	211
Hazus 6.1 Inventory	245	128	211

The second main improvement in the *Synthesized National* inventory over NSI is the assignment of the number of residential units in multifamily buildings through its use of census block data. The *NSI Point* and *NSI Spatial Join* inventories provide only ranges of the number of residential units (rather than single values) and overestimate the number of RES3F structures compared to local data. The census unit scaling method introduced earlier and formalized in Equation 1 addresses both issues. Figure 14 shows the correlation coefficient between the number of

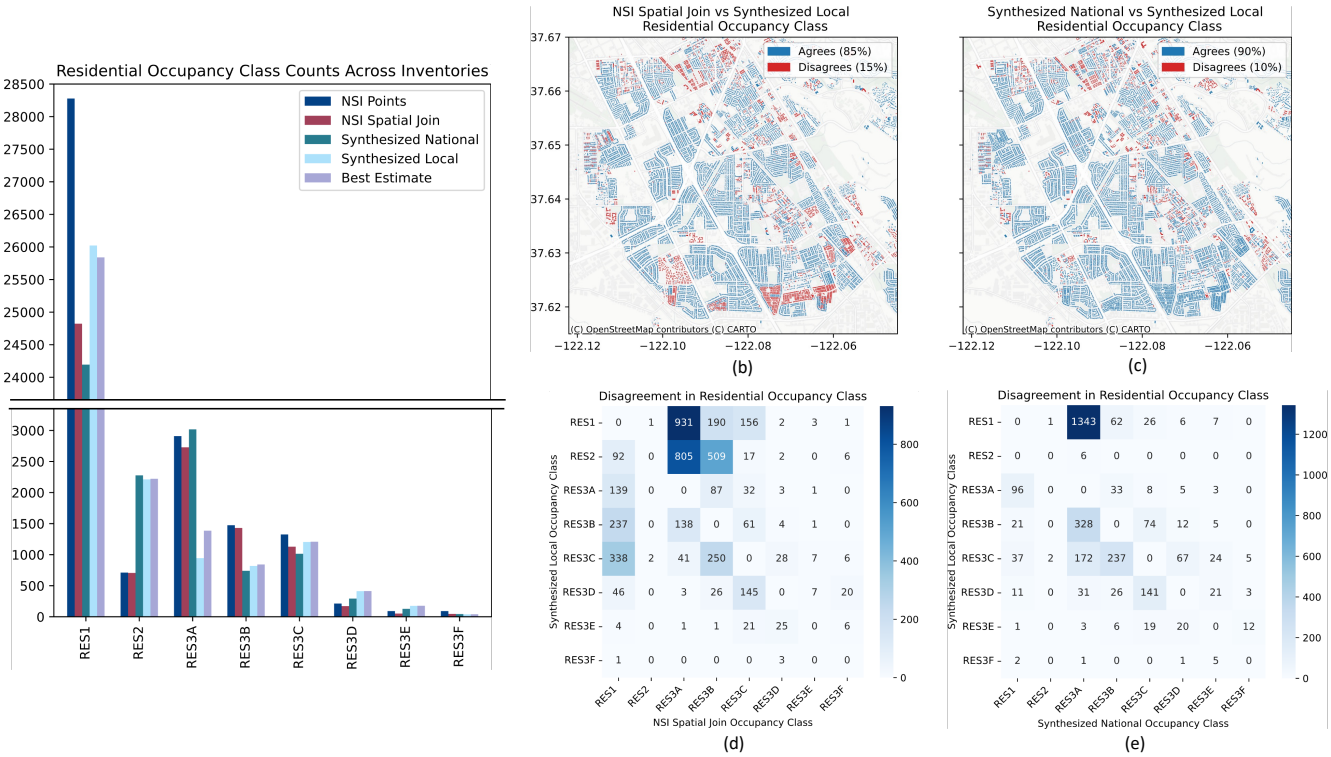

Figure 13: (a) Building counts by residential occupancy class in each inventory. Footprint-level agreement of residential occupancy class comparing the *Synthesized Local* inventory with the (b) *NSI Spatial Join* and (c) *Synthesized National* inventories. Confusion matrices for disagreeing footprints comparing the *Synthesized Local* inventory with the (d) *NSI Spatial Join* and (e) *Synthesized National* inventories.

units in the *Best Estimate* inventory (corresponding to the number of address points attributed to a given footprint, which is the most robust estimation method given local data) and the number of units estimated using various approaches: Equation 1 (14a), the average of the NSI unit range (14b), the minimum of the NSI unit range (14c), and the maximum of the NSI unit range (14d). The Equation 1 census unit estimation method (14a) shows the strongest correlation with the *Best Estimate* Inventory across the investigated methods for the *Synthesized National* Inventory. In addition, Figure 14 shows the total number of RES3F structures estimated using Equation 1 (14a) provides the closest approximation of the true number of RES3F buildings. The census housing unit scaling method overestimates RES3A structures compared to the local inventory because RES3A serves as a lower bound for unit assignment in multifamily footprints. This overestimation has minimal impact on the risk assessment since the structure type assignment for RES1 and RES3A are identical in this study.

Although aggregated occupancy class patterns across inventories align reasonably well, this is not true for all building features. As summarized in Figure 15a, classifications for building type (W, H, M, C, S), which describes primary building material and can be used to infer structure type, have significant discrepancies across inventories, even at an aggregated level. Tax parcels indicate that most structures in Hayward are wood frames, yet the *NSI Point* and *NSI Spatial Join* inventories classify a large percentage (about 40%) as masonry, compared to just 4% in the *Synthesized Local* inventory. Unlike occupancy class, as illustrated in Figure 15b, building type disagreements do not appear geospatially clustered.

To improve building type assignment in the *Synthesized National* inventory, the structure types are sampled for each footprint according to the Hazus building stock mapping scheme tables based on occupancy class, year built, and number of stories, without considering the NSI-assigned building type [66]. Based on Figure 15a, building type designation from NSI is biased, greatly overestimating the presence of masonry; thus, including it in the structure type sampling would result in biased selection of vulnerability models. If structure types are sampled without regard to NSI building type, building type can be extracted back from the selected structure type (i.e., a footprint with selected structure type C1 would have building type C). For commercial and industrial buildings, sampling structure type without considering NSI building type does not significantly improve building type classification; however, it notably improves residential classifications, correctly identifying most 1-4 unit homes as wood framed. This can be observed in Figure 15c. While the commercial and industrial footprints on the city's west side still see

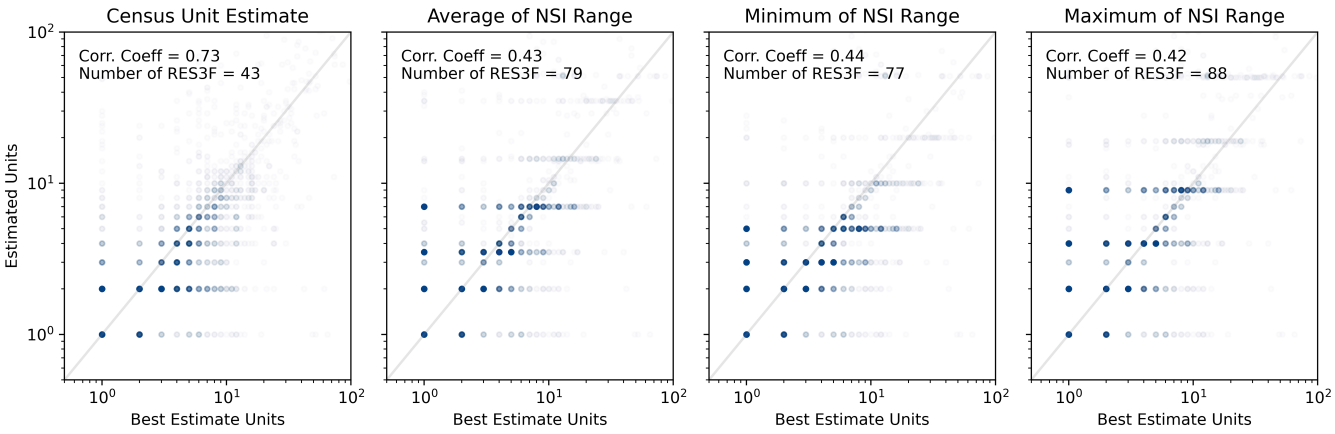

Figure 14: Comparison of housing unit estimation methods from the national workflow, using housing units from the *Best Estimate* inventory as reference. Each plot displays the corresponding correlation coefficient and the number of 50+ unit multi-family (RES3F) buildings. There are 40 RES3F buildings in the *Best Estimate* inventory.

high levels of building type disagreement with the *Synthesized Local* inventory, the majority of smaller residential footprints making up large portions of the city now agree with the *Synthesized Local* data. Local data can be beneficial for specifying building type at the footprint level, which can in turn be used to sample appropriate structure types. However, if local data is not available, it is more effective to disregard NSI's building type labels due to their overestimation of masonry structures.

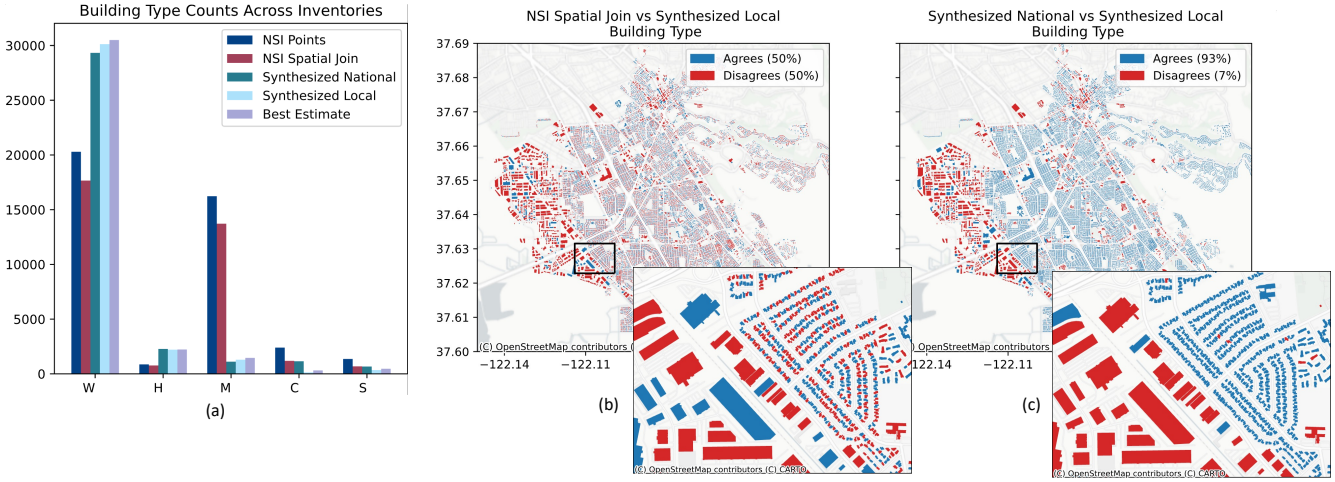

Figure 15: (a) Building type count by inventory. Footprint-level building type agreement between (b) *NSI Spatial Join* and *Synthesized Local* and (c) *Synthesized National* and *Synthesized Local*.

The comparisons of occupancy class and building type demonstrate that carefully attributing data into footprints and synthesizing multiple data sources, even when relying only on national datasets, can produce a robust inventory that more faithfully represents local conditions. Table 6 summarizes the percentage agreement across different building inventories, demonstrating the benefits of the *Synthesized National* inventory. In terms of occupancy class, RES1 footprints show high agreement across all sources; however, comparisons amoung buildings other than single-family homes (RES1) make the benefits of the *Synthesized National* inventory even more apparent.

6. Summary and Conclusions

This study outlines important considerations and proposes a systematic framework for synthesizing multiple data sources to develop a building inventory. The framework can accommodate various input data types, and it can be used to develop inventories of various fidelities, as characterized by spatial resolution, typology resolution, and accuracy. The framework involves preprocessing data, attributing sources to a baseline geometry, selecting values for features with disagreement, filling gaps in feature values through imputation or inference, and mapping

Table 6: Footprint-Level Building Type and Occupancy Class Comparisons for Synthesized Local Inventory vs NSI Spatial Join and Synthesized National Inventories.

	Percent Agreement between Spatial Join and Synthesized Local	Percent Agreement between Synthesized National and Synthesized Local
Building Type: All Footprints	50%	93%
Occupancy Class Category: All Footprints	97%	97%
Occupancy Class: All Footprints	81%	86%
Occupancy Class: All Footprints, Excluding RES1	38%	59%
Occupancy Class: All Residential Footprints, Excluding RES 1	46%	75%

to features required for simulation. The method is applied to a case study of Hayward, CA to illustrate the inventory development framework and evaluate how the choice of data sources and inventory development methods change the resulting inventory. Specific methods were developed to attribute building point data and parcel data to building footprints, synthesize NSI and HIFLD datasets, and leverage census data to improve the estimation of the number of housing units using national data. Comparisons of results for five building inventories, all nominally representing Hayward circa 2023, demonstrate the significant differences that can arise in building features such as occupancy class, number of units, and building type. A parallel research effort is applying these varying inventories to directly quantify their impact on seismic risk estimates at the city scale.

The following takeaways from the case study presented in this paper are expected to be applicable to inventory development broadly within the United States.

Different data sources and data synthesis methods lead to significant differences in the resulting building inventories. This study revealed significant differences in the quantity and distribution of building features such as occupancy class and building type across the city of Hayward, depending on the choice of input data (national, local, etc.) and inventory synthesis methods. Aggregate metrics, such as total building count or population, only vary moderately across inventories; however, more substantial differences appear when the looking at more detailed features, such as occupancy class or building type. To help contextualize and make regional-level assessments more usable for risk mitigation, more standardized and systematic approaches towards inventory development and quantification of sensitivity to inventory fidelity are needed.

Disagreements between sources do not occur at random, exacerbating biases. Disagreements between national data sources like NSI and local data do not occur at random. The observed disagreements in this study are clustered both geospatially and by building features (e.g., occupancy classes), which introduced additional biases in the results. For example, NSI greatly overestimates the presence of masonry in the residential building inventory when compared to local data sources. Similarly, the NSI and Hazus 6.1 inventories in Hayward significantly underestimate the number of mobile and manufactured homes compared to tax data, HIFLD data, and a manual evaluation. These homes are seismically vulnerable and often house socioeconomically vulnerable individuals, illustrating how errors in the inventory and the corresponding biases in results can lead to studies missing significant risk factors for a community.

Imperfect data can be improved by systematically synthesizing multiple data sources into a single inventory. This paper presents a systematic framework to synthesize information from multiple data sources with different source geometries, including footprints, parcels, points, and multi-building polygons such as census blocks and zoning data. Synthesizing multiple data sources carefully leverages the strengths of each, including the comprehensive coverage of national inventories like NSI, the accuracy of occupancy-class specific data from HIFLD, and the local specificity of tax and address data. This synthesis can also help identify systematic inaccuracies by highlighting between-source disagreements.

Synthesizing nationally available and locally obtained data leads to the best estimate inventory; however, if only national datasets are used, there are simple steps that can help reduce bias and address common issues. This study develops a "best estimate inventory" by synthesizing national and local inputs. Whereas standardized methods can be developed for common national datasets, integration of local tax and address data requires considerable effort because specific processes need to be developed to obtain, preprocess, and synthesize data that follows the specific schema used by each local jurisdiction. The national synthesis workflow, summarized in Figure 4, presents a straightforward way to improve on NSI, which is a commonly used national building inventory. The workflow addresses some common errors and biases observed in the Hayward, which likely exist in other locations in NSI data. These include 1) approximate geometries suggesting structures where none exist, 2) misrepresenting multi-family buildings as separate single-family points located in a single footprint, 3) significantly under-representing mobile and manufactured housing, 4) lacking specific unit counts and overestimating

the presence of 50+ unit buildings, and 5) inaccurately representing educational and emergency response structures. By synthesizing multiple nationally available data sources, the resulting inventory can provide a much better approximation of local data.

There are several inventory development concepts that this study does not explicitly address that could benefit from future work. The first is the temporality of the inventory. This study combines data from different years, so the resulting synthesized inventories do not strictly represent the city at a specific year, unlike for example, population censuses based on data collected at specific times. For the Hayward example, all data sources originate from years between 2018 and 2025 (see source citations in Table 1). Since Hayward is not a particularly fast-growing city, the benefits gained by synthesizing across different data sources outweighed the drawbacks of mixing reference years. This tradeoff may be different in other contexts where the built environment is changing due to rapid urbanization. Population data and the associated risk to people may also exhibit seasonal shifts due to university students, tourism, or other factors.

Future work should also quantify inventory uncertainty. This study resolves within- and between-source disagreements and fills gaps by selecting single instances of building feature values. While this approach is practical, synthesizing multiple data sources in this way can obscure varying levels of uncertainty in various inventory components and attributes. A more robust approach would be to capture and represent uncertainties in the inventory data, which can then be propagated through a risk assessment. Finally, it is important to note that this discussion focuses on data-rich contexts such as the United States, where multiple datasets can be leveraged and combined to produce a footprint-level inventory. Although the framework and methods discussed throughout the study can be adapted for use elsewhere, fundamentally different approaches might be warranted in data-scarce environments. Studies from GEM and others recommend methods they found useful to create building inventories in such contexts [19, 22, 68].

In summary, building inventories are essential for representing the physical assets, social organizations, and populations that are exposed to natural hazards in regional risk assessments. Collectively developing and improving shared methods for synthesizing inventory data from multiple sources is proposed as an effective way to identify potential biases and resolve disagreements in the data to more faithfully capture the built environment. The framework and methods presented in this paper are provided to help facilitate systematic synthesis tools for developing high-fidelity inventories.

Software and Data Availability

All Python functions, example scripts, data, and building inventories developed for this study are available on GitHub at `https://github.com/mlochhead/Building_Inventory_Generation`. The core Python functions used to generate the building inventories can be applied to other case studies beyond the one presented here. Within the repository, the *hayward_ example* folder contains all input data, Jupyter notebooks for generating the Hayward inventories, R2D input files, and results.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program, a Stanford Graduate Fellowship, and the SimCenter (supported by the National Science Foundation under Grants CMMI-1612843 and 2131111). Any options, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank Dr. Jack Baker, Dr. David Lallemant, and Doug Bausch for their support and feedback throughout the project, as well as their expertise in building inventories and regional risk assessment. The authors would also like to thank AnnaElisa Huynh for her review and feedback.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to edit some sections of the manuscript for language and clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Study Terminology

Accuracy: The degree to which the modeled building inventory correctly reflects real-world building characteristics, regardless of the level of spatial or typology resolution.

Baseline Geometry: The selected spatial representation used to represent the final inventory. Multiple data sources are synthesized by attributing each to the single baseline geometry.

Between-source disagreement: When two different data sources provide conflicting information about the features of the same building footprint.

Bounding Geometry: Polygon geometry used to assist the process of attributing data to footprints. The bounding geometry polygons should be exhaustive (cover the space completely).

Building Features: The descriptive characteristics of a building provided by a data source, excluding its geometry. These include physical attributes (e.g., number of stories, structural system), use-related information (e.g., residential, commercial), and socio-economic details (e.g., number of residents, income, tenure).

Fidelity: The extent to which the inventory accurately and comprehensively represents the real building stock. Creating higher-fidelity inventories involves increasing the spatial and/or typology resolution, while maintaining at least as much accuracy as the lower-resolution inventory.

Imputation: The process of filling in missing values using statistical methods based on patterns found in existing data.

Inference: The process of estimating missing building features by applying logical assumptions or using information from external sources.

Source Geometry: The spatial representation used by a specific data source to represent building location or extent. Source geometry vaies based on what the data source describes. Examples include building footprint polygons, points at building centroids, or aggregated geometries such as census blocks (which describe multiple buildings).

Spatial Resolution: The geographic unit at which the inventory information is provided. For example, data could be provided at the individual building level (higher spatial resolution) or aggregated at the census tract level (lower spatial resolution).

Typology Resolution: The level of detail used to describe the building itself. For example, a broad category like "light frame wood construction" represents a lower typology resolution, whereas a more specific description, such as "2-story wood house from 1962 with an elevated crawlspace," represents a higher typology resolution.

Within-source disagreement: When multiple records from the same data source are linked to one building footprint but report conflicting information about its features.

Appendix B. Building Feature Terminology and Source

Building Type: This study uses the term *building type* to describe the primary structural material. This term is adopted from NSI, which states that this field (with five material options) is used for structure stability functions. The term building type is also used in the Hazus 6.1 Inventory, and the use case in this study matches closest with the five general building types used in Hazus (wood frame, steel frame, concrete, masonry, and manufactured housing). Hazus also uses the term "specific building type" (SBT); however, the SBT definitions are more specific than those adopted for building type in this study.

Occupancy Class: This study adopts the term *occupancy class* to describe the primary purpose of a building, for example "single-family dwelling" or "hospital." This term is adopted from the Hazus inventory. NSI uses a similar "Occupancy Type" field to determine structure valuation, population, and to define structure damage criteria. More details can be found below. The possible values for occupancy class are adopted from Hazus, with some extensions. Changes to the list of possible occupancy classes are described in Appendix C.

Structure Type: is study uses the term *structure type* to describe the primary structural system. Structure types are more specific than building types (i.e. if the building type is 'steel', the structure type could be 'steel moment frame', 'steel braced frame,' etc.). The term structure type, as used in this study, maps closely to the specific building types (SBTs) for earthquakes, as specified in Table 4.2 of the Hazus 6.1 Inventory Manual.

Appendix C. Extended Occupancy Class Information

This study adopts the term *occupancy class* to describe the primary purpose of a building, for example "single-family dwelling" or "hospital." This term is adopted from the Hazus inventory. Most of the possible values for occupancy class are adopted directly from Hazus, with some extensions. The Hazus occupancy class options adopted directly are shown below (Table 4-1 in the Hazus 6.1 Inventory Manual).

Table 4-1 Hazus General and Specific Occupancy Classes

Hazus General Occupancy Class	Hazus Specific Occupancy Class	Class Description
Residential	RES1	Single-family Dwelling
Residential	RES2	Mobile Home
Residential	RES3A	Multi-Family Dwelling – Duplex
Residential	RES3B	Multi-Family Dwelling – 3-4 Units
Residential	RES3C	Multi-Family Dwelling – 5-9 Units
Residential	RES3D	Multi-Family Dwelling – 10-19 Units
Residential	RES3E	Multi-Family Dwelling – 20-49 Units
Residential	RES3F	Multi-Family Dwelling – 50+ Units
Residential	RES4	Temporary Lodging
Residential	RES5	Institutional Dormitory
Residential	RES6	Nursing Home
Commercial	COM1	Retail Trade
Commercial	COM2	Wholesale Trade
Commercial	COM3	Personal and Repair Services
Commercial	COM4	Business/Professional/Technical Services
Commercial	COM5	Depository Institutions (Banks)
Commercial	COM6	Hospital
Commercial	COM7	Medical Office/Clinic
Commercial	COM8	Entertainment & Recreation
Commercial	COM9	Theaters
Commercial	COM10	Parking
Industrial	IND1	Heavy
Industrial	IND2	Light
Industrial	IND3	Food/Drugs/Chemicals
Industrial	IND4	Metals/Minerals Processing
Industrial	IND5	High Technology
Industrial	IND6	Construction
Agriculture	AGR1	Agriculture
Religion	REL1	Church/Non-Profit
Government	GOV1	General Services
Government	GOV2	Emergency Response
Education	EDU1	Schools/Libraries
Education	EDU2	Colleges/Universities

Figure C.16: Hazus General and Specific Occupancy Classes (Table 4-1 from Hazus 6.1 Inventory Manual)

In addition to the above list, several possible occupancy class values were created to extend beyond the existing list for various reasons. The occupancy class extensions are shown below.

Table C.7: Occupancy class values in this study extended off of the Hazus occupancy class values

New Occupancy Class	**Reason for Adding Occupancy Class**	**Description of Occupancy Class**
EDU1-PRIV	HIFLD more specific than Hazus Occupancies	This is used when occupancy class is assigned to HIFLD data from the Private School dataset
EDU1-PUB	HIFLD more specific than Hazus Occupancies	This is used when occupancy class is assigned to HIFLD data from the Public School dataset
GOV2-POLICE	HIFLD more specific than Hazus Occupancies	This is used when occupancy class is assigned to HIFLD data from the Local Law Enforcement Locations dataset
GOV2-FIRE	HIFLD more specific than Hazus Occupancies	This is used when occupancy class is assigned to HIFLD data from the Fire and Emergency Medical Service Stations dataset
GOV2-OPERATIONS	HIFLD more specific than Hazus Occupancies	This is used when occupancy class is assigned to HIFLD data from the State or Local Emergency Operations Centers
RES3	Lack of specificity in local data descriptions	This is used when data specifies that a footprint is multi-unit residential, but no information is provided on the number of units. An example from the Hayward Address Point Data is: 'Multi-Family Dwelling'
IND	Lack of specificity in local data descriptions	This is used when data specifies that a footprint is industrial, but there is insufficient specificity for a more detailed industrial designation. An example from the Hayward Address Point Data is: 'Industrial Facility'
COM	Lack of specificity in local data descriptions	This is used when data specifies that a footprint is commercial, but there is insufficient specificity for a more detailed commercial designation. An example from the Hayward Parcel Data is: 'Miscellaneous Improved Commercial'
NOTBLDG	Local data corresponds to something that is not a building	This is used when a tax parcel or address point has a description that implies that the parcel is not a building. Examples from the Hayward Parcel Data are: 'Traffic Signal,' 'Telecom Utility Box', and 'Golf Course'
UNK	Lack of specificity in local data descriptions	This is used when there is insufficient specificity for a more detailed occupancy class designation. An example from the Hayward Address Point Data is: 'Building General'
Suffix: _VAC Examples: RES3_VAC COM_VAC	Local data specifies "planned" or "vacant"	_VAC is appended to the end of an existing occupancy class when the data implies that the specified point or parcel is either vacant or planned. If multiple data sources confirm a _VAC description and there is no year built available, it is assumed that there is no building present. Examples from the Hayward Parcel Data include: 'Vacant apartment land, capable of 5 or more units' (gets mapped to RES3_VAC), or 'Townhouse – Planned Development' (gets mapped to RES3_VAC)
Suffix: M Examples: RES3AM	No existing designation for mixed use occupancy classes	M is appended to the end of an existing residential occupancy class with the data implies that the footprint is mixed use. Examples from the Hayward Parcel Data include 'Multiple-Res building of 5 or more units + commercial units' (gets mapped to RES3M). The residential occupancy is maintained as the base because it is used for structure type assignment. Further benefits could be gained by having vulnerability models for mixed-use structures.

Appendix D. Data Source Details

National Data Sources

- **National Structures Inventory (NSI)**: NSI is a single-building point source containing a broad range of building features associated with the location of the building centroid. NSI is developed by synthesizing many existing data sources, including several Hazus datasets, several HIFLD datasets (Lightbox, nursing home, hospital, mobile home, and building layers), ESRI's business layer, Microsoft building footprints, FEMA's USA structures polygons, US Census data, the NCES school dataset, and USGS national elevation dataset [10]. NSI is available at two levels: public and private. The public data is open access and contains many building features, including occupancy class, number of stories, replacement cost, and more. The private data contains additional details, including the parcel number (APN), the number of housing units in a building, the reported year built, and more. Although the private data is available to the Hazus team and used in the development of Hauzs 6.1 [9], it is not broadly available to researchers. NSI is comprehensive in that it does not have any gaps or missing features; however, some building features appear to be more robust than others. Based on observations in this study, occupancy class is one of the more robust features of NSI, particularly in distinguishing between residential and non-residential areas. In the public data, the year built is only available as the median across a census block group, which can be limiting when assigning structural vulnerability, and residential unit counts are only provided as broad ranges (e.g., 5-10, 20-50). Furthermore, the dataset contains some anomalies, including a high number of points located along transit lines or far from buildings, many single-family homes stacking into a single footprint (likely representing a multi-unit building), and instances where the number of stories is unrealistically high. Only limited documentation on the development of NSI is available, and the latest release was in 2022. While comprehensive, NSI has notable limitations, especially when relying only on the publicly available fields.

- **USA Structures**: USA Structures is a single-building polygon source describing all structures larger than 450 square feet in the United States, originally developed for flood insurance, mitigation, and response purposes. The set of building footprint polygons was developed by extracting outlines via commercially available satellite imagery, leveraging machine learning techniques, and synthesizing data from multiple sources, including local governments and the National Geospatial-Intelligence Agency. The dataset is publicly available and has several building features beyond just the footprint polygon, including footprint height (somewhat sparsely available), and occupancy class. Unlike NSI, it does not have extensive building features available regarding value, structural characteristics, number of units, year built, etc. In the case study explored in this paper, USA Structures footprints were comprehensive and were only missing some newer construction. In other locations, it has been noted that this dataset does have some gaps in footprint coverage, particularly in residential areas. This dataset is actively being developed, and documentation describes that it is set up to have more features such as first floor height added in the future. The latest update was in 2025.

- **Homeland Infrastructure Foundation-Level Data (HIFLD)**: HIFLD Open Data consists of many individual datasets describing critical infrastructure across the US. Some HIFLD datasets are single-building points (e.g., fire stations) and others are multi-building polygons (e.g., college/university campuses), among other types. HIFLD was originally developed for several broad purposes, including emergency management, critical infrastructure protection, and national operations and data fusion centers. Based on observations in this study, HIFLD is comprehensive for the critical infrastructure it describes (public schools, fire stations, etc.). Thus, its scope does not aim to be comprehensive for all buildings, like NSI and USA Structures, but instead geographically represent the infrastructure within the scope of the individual dataset. Thus, it is a very valuable resource for describing specific subsets of the building inventory, such as schools, police and fire stations, emergency operations centers, and more. Beyond just the location and building type, HIFLD does carry some additional information that may be helpful for building inventory development depending on the dataset, such as school enrollment population. This dataset was actively developed and updated for many years, but the open access version has been discontinued as of August 26th, 2025. The HIFLD datasets discussed in this paper, though no longer available directly, can be obtained through the Hazus 7.0 inventory, which is available at the FEMA Flood Map Service Center [69].

Local Data Sources

- **Tax Parcel Data:** Tax parcel data may be available at the city or county level, and it is typically a mixed-type polygon source based on parcel geometries. Tax data is typically linked to an Assessor Parcel

Number (APN), and tax parcels correspond to owned area. Thus, a tax parcel geometry does not have a 1:1 relationship with a building footprint geometry, introducing further complexity in merging various datasets, which we will discuss later in the paper. Beyond the APN, the available features vary widely based on location.

- **Address Data**: Address data is typically a single-unit point source that corresponds to the locations of individual addresses, and may or may not be linked to an APN. These points do not necessarily represent a building centroid or any specific location within a building, and in some cases may not correspond to a building at all—for example, a P.O. box or a central mail location serving multiple buildings. It is therefore important to understand how address data is defined in the specific context of use.

- **Zoning Data:** Zoning data is a multi-building polygon source and represents land-use regulations assigned by local jurisdictions. These polygons define allowable uses for buildings in the area, density, and other planning constraints. However, zoning designations do not necessarily reflect the existing built environment, as many parcels may be used in ways that predate current zoning codes or are permitted under special exceptions. Furthermore, the meaning of different zoning designations is often defined at the local level, and must be interpreted accordingly.

- **Building Permit Data:** Building permit data, if available, contains records of construction, renovation, or demolition activities, and it is typically non-spatial data. Permits may include details such as construction type, year, square footage, occupancy, or number of units. However, permit records may not capture all building activity, especially for older structures, unpermitted work, or jurisdictions with inconsistent reporting. In addition, the format and availability of permit data vary widely across locations. Building permit data requires a significant amount of preprocessing to extract helpful building inventory information, though machine learning techniques may be helpful in parsing this information in the future.

Global Footprint Data Sources

- **Microsoft Building Footprints**: Microsoft has a global building footprint database (single-building polygons), generated by deploying machine learning methods on Bing Maps imagery between 2014 and 2024. This dataset does not contain any building features outside of the footprint polygon and footprint height, which is somewhat sparsely populated and is only available in some regions. While increasingly accurate, computer vision models still may inadvertently label features in the landscape as building footprints or fail to recognize some footprints. This dataset is actively being developed and was most recently updated in 2025.

- **OpenStreetMaps**: OpenStreetMap is a global building footprint database (single-building polygons) which is open-source and community-developed, thus differentiating it from the other global footprints sources. Thus, coverage varies widely by location, building footprints may or may not include associated building features beyond the polygon itself, and the data does not represent a single, specific year's building inventory.

- **Overture Maps:** Overture Maps is a foundation started in 2022 dedicated to creating easy-to-use open map data. The *buildings* dataset from Overture Maps contains single-building polygon data describing human-made structures with roofs or interior spaces that are permanently or semi-permanently in one place. The geometry specified in this dataset is the outer footprint traced from satellite/aerial imagery of a building. The Overture footprints are developed by synthesizing many individual datasets, including OpenStreetMap, Esri Community Maps, Microsoft Building Footprints, Google Open Buildings, and others [59]. The synthesis process first prioritizes community-contributed data, then "fills in" the rest with the best machine learning based data available. This dataset is actively being developed and was most recently updated in 2025.

References

[1] United Nations Office for Disaster Risk Reduction, Sendai framework for disaster risk reduction 2015–2030, Tech. rep., United Nations, Geneva, Switzerland (2015).

[2] C. Yepes-Estrada, A. Calderon, C. Costa, H. Crowley, J. Dabbeek, M. C. Hoyos, L. Martins, N. Paul, A. Rao, V. Silva, Global building exposure model for earthquake risk assessment, Earthquake Spectra 39 (4) (2023) 2212–2235. doi:10.1177/87552930231194048.

[3] C. Scawthorne, A Brief History of Seismic Risk Assessment, Springer, Berlin, Heidelberg, Publication Location, 2006, Ch. 2, pp. 5–81. doi:10.1007/978-3-540-71158-2_2.

[4] Cotality, https://www.cotality.com/our-data, accessed 2025.

[5] LightBox, https://www.lightboxre.com/data/, accessed 2025.

[6] ATTOM, https://www.attomdata.com/solutions/bulk-data-licensing/, accessed 2025.

[7] Regrid, https://app.regrid.com/store, accessed 2025.

[8] FEMA, Hazus 6.1 Earthquake Model Technical Manual, Tech. rep., Federal Emergency Management Agency (2024).

[9] FEMA, Hazus 7.0 Inventory Technical Manual, Tech. rep., Federal Emergency Management Agency (2025).

[10] U.S. Army Corps of Engineers, National Structure Inventory, https://www.hec.usace.army.mil/confluence/nsi/technicalreferences/latest/technical-documentation (2022).

[11] Oak Ridge National Laboratory, FEMA Response Geospatial Office, USA Structures, https://gis-fema.hub.arcgis.com/pages/usa-structures (2024).

[12] H. L. Yang, M. Laverdiere, T. Hauser, B. Swan, E. Schmidt, J. Moehl, A. Reith, D. Adams, B. Morris, J. McKee, M. Whitehead, M. Tuttle, A baseline structure inventory with critical attribution for the US and its territories, Scientific Data 11 (1) (2024) 502. doi:10.1038/s41597-024-03219-x.

[13] K. Jaiswal, D. Wald, K. Porter, A Global Building Inventory for Earthquake Loss Estimation and Risk Management, Earthquake Spectra 26 (3) (2010) 731–748. doi:10.1193/1.3450316.

[14] K. S. Jaiswal, M. D. Petersen, K. Rukstales, W. S. Leith, Earthquake Shaking Hazard Estimates and Exposure Changes in the Conterminous United States, Earthquake Spectra 31 (2015) S201–S220. doi:10.1193/111814EQS195M.

[15] K. S. Jaiswal, D. J. Wald, Development of a semi-empirical loss model within the USGS Prompt Assessment of Global Earthquakes for Response (PAGER) System, in: Proceedings of the 9th US and 10th Canadian Conference on Earthquake Engineering: Reaching Beyond Borders, 2010, pp. 25–29.

[16] S. A. Figueira, M. Amini, D. T. Cox, A. R. Barbosa, Methodology for Virtual Damage Assessment and First-Floor Elevation Estimation: Application to Fort Myers Beach, Florida and Hurricane Ian (2022), Natural Hazards Review 26 (2) (2025) 04025012.

[17] O. E. J. Wing, W. Lehman, P. D. Bates, C. C. Sampson, N. Quinn, A. M. Smith, J. C. Neal, J. R. Porter, C. Kousky, Inequitable patterns of US flood risk in the Anthropocene, Nature Climate Change 12 (2) (2022) 156–162. doi:10.1038/s41558-021-01265-6.

[18] B. Cetiner, F. McKenna, S.-r. Yi, B. Wang, I. V. Manousakis, BRAILS++ (2025).
URL https://github.com/NHERI-SimCenter/BrailsPlusPlus

[19] F. Dell'Acqua, P. Gamba, K. Jaiswal, Spatial aspects of building and population exposure data and their implications for global earthquake exposure modeling, Natural Hazards 68 (3) (2013) 1291–1309. doi:10.1007/s11069-012-0241-2.

[20] V. Silva, D. Amo-Oduro, A. Calderon, C. Costa, J. Dabbeek, V. Despotaki, L. Martins, M. Pagani, A. Rao, M. Simionato, D. Viganò, C. Yepes-Estrada, A. Acevedo, H. Crowley, N. Horspool, K. Jaiswal, M. Journeay, M. Pittore, Development of a global seismic risk model, Earthquake Spectra 36 (2020) 372–394. doi:10.1177/8755293019899953.

[21] H. Santa María, M. A. Hube, F. Rivera, C. Yepes-Estrada, J. A. Valcárcel, Development of national and local exposure models of residential structures in Chile, Natural Hazards 86 (1) (2017) 55–79. doi:10.1007/s11069-016-2518-3.

[22] C. Yepes-Estrada, V. Silva, J. Valcárcel, A. B. Acevedo, N. Tarque, M. A. Hube, G. Coronel, H. S. María, Modeling the Residential Building Inventory in South America for Seismic Risk Assessment, Earthquake Spectra 33 (1) (2017) 299–322. doi:10.1193/101915eqs155dp.

[23] G. G. Deierlein, F. McKenna, A. Zsarnóczay, T. Kijewski-Correa, A. Kareem, W. Elhaddad, L. Lowes, M. J. Schoettler, S. Govindjee, A Cloud-Enabled Application Framework for Simulating Regional-Scale Impacts of Natural Hazards on the Built Environment, Frontiers in Built Environment 6 (Nov. 2020). doi:10.3389/fbuil.2020.558706.

[24] J. W. van de Lindt, J. Kruse, D. T. Cox, P. Gardoni, J. S. Lee, J. Padgett, T. P. McAllister, A. Barbosa, H. Cutler, S. Van Zandt, N. Rosenheim, C. M. Navarro, E. Sutley, S. Hamideh, The interdependent networked community resilience modeling environment (IN-CORE), Resilient Cities and Structures 2 (2) (2023) 57–66. doi:10.1016/j.rcns.2023.07.004.

[25] E. M. Rathje, C. Dawson, J. E. Padgett, J.-P. Pinelli, D. Stanzione, A. Adair, P. Arduino, S. J. Brandenberg, T. Cockerill, C. Dey, M. Esteva, F. L. Haan, M. Hanlon, A. Kareem, L. Lowes, S. Mock, G. Mosqueda, DesignSafe: New Cyberinfrastructure for Natural Hazards Engineering, Natural Hazards Review 18 (3) (2017) 06017001. doi:10.1061/(ASCE)NH.1527-6996.0000246.

[26] L. Dahal, H. Burton, K. Zhong, High-Fidelity High-Resolution Regional Seismic Risk and Resilience Assessment of Large Building Inventories, Earthquake Engineering & Structural Dynamics 54 (5) (2025) 1376–1396. doi:10.1002/eqe.4313.

[27] L. Ceferino, J. Mitrani-Reiser, A. Kiremidjian, G. Deierlein, C. Bambarén, Effective plans for hospital system response to earthquake emergencies, Nature Communications 11 (1) (2020) 4325. doi:10.1038/s41467-020-18072-w.

[28] T. Bassman, A. Zsarnóczay, J. Saw, S. Wang, G. Deierlein, High-Fidelity Testbed Development for Regional Risk Assessment in Alameda, California, in: 12th National Conference on Earthquake Engineering, 2022.

[29] M. Hilt, Early Results from a Methodology to Leverage Seismic Risk Assessments to Inform Seismic Policy Development in the City of Vancouver, Tech. Rep. 8918 (2022). doi:10.4095/330927.

[30] A. Zsarnoczay, G. G. Deierlein, F. McKenna, M. Schoettler, S.-r. Yi, B. Cetiner, A. B. Satish, J. Zhao, J. Bonus, A. F. Melaku, et al., An open-source simulation platform to support and foster research collaboration in natural hazards engineering, Frontiers in Built Environment 11 (2025). doi:https://doi.org/10.3389/fbuil.2025.1590479.

[31] J. Dabbeek, H. Crowley, V. Silva, G. Weatherill, N. Paul, C. I. Nievas, Impact of exposure spatial resolution on seismic loss estimates in regional portfolios, Bulletin of Earthquake Engineering 19 (14) (2021) 5819–5841.

[32] M. Erdik, Earthquake risk assessment, Bulletin of Earthquake Engineering 15 (12) (2017) 5055–5092. doi:10.1007/s10518-017-0235-2.

[33] G. Pavic, M. Hadzima-Nyarko, B. Bulajic, Z. Jurkovic, Development of Seismic Vulnerability and Exposure Models—A Case Study of Croatia, Sustainability 12 (3) (2020) 973. doi:10.3390/su12030973.

[34] T. Anagnos, M. Comerio, C. Goulet, J. Steele, J. Stewart, Development of a concrete building inventory: Los Angeles case study for the analysis of collapse risk, in: Proc. 9th National & 10th Canadian Conf. on Eq. Eng, 2010.

[35] Applied Technology Council, San Francisco Tall Buildings Study, Tech. Rep. ATC-119-1, City and County of San Francisco (2018).

[36] A. Djenaliev, M. Kada, A. Chymyrov, Building inventory data development for pre-earthquake evaluation., International Journal of Geoinformatics 12 (4) (2016).

[37] D. M. Wiebe, D. T. Cox, Application of fragility curves to estimate building damage and economic loss at a community scale: a case study of Seaside, Oregon, Natural Hazards 71 (3) (2014) 2043–2061. doi:10.1007/s11069-013-0995-1.

[38] W. Nurkarim, A. W. Wijayanto, Building footprint extraction and counting on very high-resolution satellite imagery using object detection deep learning framework, Earth Science Informatics 16 (2022) 515–532.

[39] Z. Li, Q. Xin, Y. Sun, M. Cao, A Deep Learning-Based Framework for Automated Extraction of Building Footprint Polygons from Very High-Resolution Aerial Imagery, Remote Sensing 18 (2021).

[40] R. Kalfarisi, M. Hmosze, Z. Y. Wu, Detecting and geolocating city-scale soft-story buildings by deep machine learning for urban seismic resilience, Natural Hazards Review 23 (1) (2022).

[41] Q. Yu, C. Wang, F. McKenna, S. X. Yu, E. Taciroglu, B. Cetiner, K. H. Law, Rapid visual screening of soft-story buildings from street view images using deep learning classification, Earthquake Engineering and Engineering Vibration 19 (2020) 827–838.

[42] D. Gonzalez, D. Rueda-Plata, A. B. Acevedo, J. C. Duque, R. Ramos-Pollán, A. Betancourt, S. García, Automatic detection of building typology using deep learning methods on street level images, Building and Environment 177 (2020).

[43] F. Ghione, S. Mæland, A. Meslem, V. Oye, Building stock classification using machine learning: A case study for Oslo, Norway, Frontiers in Earth Science 10 (2022).

[44] T. Kijewski-Correa, B. Cetiner, K. Zhong, C. Wang, A. Zsarnoczay, Y. Guo, M. Lochhead, F. McKenna, Validation of an Augmented Parcel Approach for Hurricane Regional Loss Assessments, Natural Hazards Review 24 (3) (2023) 04023022. doi:10.1061/NHREFO.NHENG-1649.

[45] K. Angeles, T. Kijewski-Correa, Advancing parcel-level hurricane regional loss assessments using open data and the regional resilience determination tool, International Journal of Disaster Risk Reduction 95 (2023) 103818. doi:10.1016/j.ijdrr.2023.103818.

[46] G. Tocchi, M. Polese, M. Di Ludovico, A. Prota, Regional based exposure models to account for local building typologies, Bulletin of Earthquake Engineering 20 (1) (2022) 193–228. doi:10.1007/s10518-021-01242-6.

[47] B. R. Ellingwood, C. , Harvey, G. , Paolo, P. , Walter Gillis, v. d. L. , John W., , N. Wang, The Centerville Virtual Community: a fully integrated decision model of interacting physical and social infrastructure systems, Sustainable and Resilient Infrastructure 1 (3-4) (2016) 95–107. doi:10.1080/23789689.2016.1255000.

[48] M. Roohi, J. W. Van De Lindt, N. Rosenheim, Y. Hu, H. Cutler, Implication of building inventory accuracy on physical and socio-economic resilience metrics for informed decision-making in natural hazards, Structure and Infrastructure Engineering 17 (4) (2021) 534–554. doi:10.1080/15732479.2020.1845753.

[49] D. Sanderson, D. Cox, Comparison of national and local building inventories for damage and loss modeling of seismic and tsunami hazards: From parcel-to city-scale, International Journal of Disaster Risk Reduction 93 (2023) 103755. doi:10.1016/j.ijdrr.2023.103755.

[50] R. Fayjaloun, C. Negulescu, A. Roulle, P. Gehl, S. Auclair, M. Faravelli, Sensitivity of earthquake damage estimation to the input data: Case study in the Luchon valley, France, in: 3rd European Conference on Earthquake Engineering & Seismology, Bucarest, Romania, 2022.

[51] A. Calderon, C. Yepes-Estrada, V. Silva, Urban seismic risk assessment for the cities of Quito, Cali, and Santiago de los Caballeros, Executive Summary, Global Earthquake Model Foundation (2022).

[52] H. Crowley, V. Despotaki, D. Rodrigues, V. Silva, D. Toma-Danila, E. Riga, A. Karatzetzou, S. Fotopoulou, Z. Zugic, L. Sousa, S. Ozcebe, P. Gamba, Exposure model for European seismic risk assessment, Earthquake Spectra 36 (2020) 252–273. doi:10.1177/8755293020919429.

[53] R. Rincon, J. E. Padgett, Fragility modeling practices and their implications on risk and resilience analysis: From the structure to the network scale, Earthquake Spectra 40 (1) (2024) 647–673. doi:10.1177/87552930231219220.

[54] J. Zou, D. Welch, A. Zsarnoczay, A. Taflanidis, G. Deierlein, Surrogate modeling for the seismic response estimation of residential wood frame structures, in: Proceedings of the 17th world conference on earthquake engineering, Japan, 2020.

[55] FEMA, Hazus 6.1 Inventory Technical Manual, Tech. rep., Federal Emergency Management Agency (2024).

[56] Geospatial Management Office, U.S. Department of Homeland Security, Homeland Infrastructure Foundation-Level Data, https://hifld-geoplatform.hub.arcgis.com/ (Accessed 2025).

[57] Microsoft, US Building Footprints, https://github.com/microsoft/USBuildingFootprints, accessed: 2025.

[58] OpenStreetMap contributors, Building Footprints, https://www.openstreetmap.org, accessed: 2025.

[59] Overture Maps Foundation, Overture buildings, v1.9.0, https://docs.overturemaps.org/guides/buildings/#14/32.58453/-117.05154/0/60, accessed: 2025.

[60] Zillow, https://www.zillow.com/research/data/, accessed 2025.

[61] NASA, Landsat Science, https://landsat.gsfc.nasa.gov/, accessed 2025.

[62] C. D. S. Ecosystem, Sentinel 2, https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-data/sentinel-2, accessed 2025 (2025).

[63] Google Maps, Street View Static API Overview, https://developers.google.com/maps/documentation/streetview/overview.

[64] Mapillary, https://www.mapillary.com/, accessed 2025.

[65] City of Hayward, Hayward Open Data, https://opendata.hayward-ca.gov/, accessed: 2025.

[66] FEMA, Hazus 6.0 Inventory Technical Manual, Tech. rep., Federal Emergency Management Agency (2022).

[67] California Seismic Safety Commission, Status of the unreinforced masonry building law, Report to the Legislature SSC 2005-02, California Seismic Safety Commission, Sacramento, CA (2005).

[68] D. Felsenstein, E. Elbaum, T. Levi, R. Calvo, Post-processing HAZUS earthquake damage and loss assessments for individual buildings, Natural Hazards 105 (1) (2021) 21–45. doi:10.1007/s11069-020-04293-1.

[69] FEMA, Flood Map Service Center: Hazus, https://msc.fema.gov/portal/resources/hazus, accessed 2025.