Exposure Matters: A Synthesis Framework for High-Resolution Building Inventory Development

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

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

Abstract

Regional simulations of the impact of natural hazards on the built environment can be used to support disaster risk management and guide mitigation priorities, and 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 higher-fidelity 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 data sources; however, few have assessed the impact of these inventory development decisions or the variability of the resulting inventory. 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 and synthesis methods can lead to substantial differences in the resulting inventory and measured risk in a community. Furthermore, these differences are geospatially clustered and concentrate in certain types of buildings, which can lead to significant biases in the results. This paper aims to both motivate and address 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, 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 non-governmental organizations. Detailed building inventories were first developed in the United 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 data, 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. One of 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 United States [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 1) the National Structures Inventory (NSI) by the US Army Corps of Engineers [10], 2) USA Structures by a collaboration between the Oak Ridge National Laboratory, USGS, and FEMA's Response Geospatial Office [11, 12], and 3) the global building inventory by the USGS for their Prompt Assessment of Global Earthquakes for Response (PAGER) service [13]. These data sources have been widely adopted for emergency preparedness, flood insurance, rapid earthquake impact estimation, and other regional studies [14, 15, 16, 17, 18]. Finally, non-governmental organizations, particularly the Global Earthquake Model (GEM), have also made substantial contributions to global inventory development [19, 20, 21, 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 data sources, and employing both top-down and bottom-up methodologies [2, 22, 23].

Recent studies demonstrate the potential of high-fidelity regional models at the level of a city, county, or urban area to support disaster risk management, inform policy decisions, and prioritize mitigation strategies [24, 25, 26, 27, 28, 29, 30]. 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 [31]. However, many of the aforementioned data sources, 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 data sources are designed for risk assessment and effective for determining total 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 [32]. The aforementioned inventories that do provide building-specific information (e.g., NSI) are not without limitations; they may contain gaps with missing buildings or features, or have filled all gaps by statistically inferring or imputing data, limiting local specificity and accuracy. This poses a challenge because data sources do not indicate which buildings or features are based on underlying data and which are modeled, making 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.

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 [33]. 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 [34, 22, 30, 35, 36, 37]. 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 data to develop inventories, particularly as more municipalities transfer their records to GIS-based data repositories [3, 38, 29]. Others leverage portions of national data sources such as the census or national inventories like NSI due to their complete 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 [39, 40], infer building features [41, 42], and classify building typologies [43, 44]. When detailed data is not otherwise available, researchers have used a variety of strategies to overcome data limitations, including simplifying assumptions, heuristic rulesets, and imputation to fill in gaps [45, 46, 27, 18]. Due to limitations in individual data sources, several studies have developed building inventories by synthesizing multiple data sources, thus producing a more comprehensive inventory [45, 35, 47]. Finally, some researchers have opted to develop synthetic inventories to bypass data availability challenges [48].

Despite the consensus on the importance of detailed inventories and the wide variety of available modeling approaches, many studies recognize significant challenges in inventory development, particularly at the building level. 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 complete, may not have the local specificity or building-specific accuracy targeted in a study. Additional challenges arise when studies require consistent inventories across multiple regions or countries with different underlying data availability (harmonized building inventories), where heterogeneous input data can make methodological consistency difficult to achieve [49, 50]. Although harmonization is most commonly discussed in the context of cross-border risk assessment, similar challenges emerge in smaller-scale case studies that draw on local data from multiple cities or counties with different data

availability.

While studies in the literature utilize various inventory data sources and data synthesis methods, few have systematically evaluated how their decisions affect building-level inventory quality or regional risk assessment outcomes beyond aggregate loss, 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) [51] 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. [51] find that reducing inventory accuracy, simulated by artificially removing data at random from a high-quality pre-existing inventory database, leads to decreased reliability in estimating damage, functionality, and displacement. Sanderson and Cox (2023) [52] 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 [52] 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. Overall results demonstrate that the NSI dataset under-predicts aggregate loss relative to the other data source [52]. A separate case study in Luchon, France also explores the impact of inventory resolution (national versus local data) on the estimation of seismic damage and concludes that a major source of uncertainty in damage estimation is the "intrinsically difficult" inventory problem [53].

These two 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 [51, 52, 53]. 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 [54]. The results demonstrate the importance of incorporating detailed information and local knowledge in earthquake risk assessments because the estimated impact (e.g., expected loss or number of collapsed buildings) can vary by an order of magnitude. Huyck et al. (2022) explore varying the level of detail used to create exposure data in Los Angeles County, revealing that building exposure databases created from more detailed data do not necessarily converge or yield more accurate loss estimates. These findings suggest that the relationship between data detail and model accuracy is not straightforward and that exposure data is an important source of uncertainty in risk modeling, in this case resulting in a factor of 2.5 on the final loss estimates [55]. Others, while not always propagating uncertainty explicitly, note both aleatory and epistemic uncertainty as important factors and current limitations of inventory development [56, 57]. Recent work has also demonstrated the consequences of such limitations in applied settings, finding that unrefined national inventories can misclassify community risk priorities, that simple targeted refinements can meaningfully improve estimates, and that the widespread practice of conducting risk assessments without exposure quality assurance can systematically mislead risk estimates and resource allocation decisions [58].

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 or have assessed how data source selection and synthesis methods affect exposure accuracy. As a result, there is a lack of standardized workflows or guidance for researchers seeking to leverage multiple data sources to develop building-level inventories. To address these gaps, this study presents 1) a systematic framework for creating footprint-level building inventories through the synthesis of multiple data sources, 2) several specific methods for implementing the framework, and 3) a quantitative evaluation of how inventory development decisions impact the resulting inventory makeup and quality. Section 2 first discusses several concepts relevant for inventory development and introduces corresponding terminology used throughout the study. Section 3 then introduces a general framework outlining steps for inventory data synthesis that are generally applicable across different contexts. Section 4 elaborates on this framework by outlining specific methods and examples of inventory synthesis in the context of seismic risk in Hayward, California, with implications shown in Section 5. Finally, Section 6 discusses how the general inventory synthesis framework can be applied to other hazard and location contexts.

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 impact 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 contains a glossary of all relevant study 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 [31]. The increasingly detailed questions posed about regional risk require higher-resolution inventories. 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 moving towards higher levels of *spatial resolution* by shifting from aggregated census-tract-level analysis to footprint-level analysis [24]. Higher spatial resolution provides more geographic detail on the building inventory and allows for the incorporation of more local hazard effects [32]. 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 [59, 60]. These two metrics of resolution can change independently from each other.

Improved *accuracy* of inventories, or the degree to which the inventory data 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 building level. Increasing the typology resolution presents even greater challenges as it demands the collection or inference of additional information. For example, while an inventory may correctly label a building as a single-family house, determining accurate information on the year built or number of stories can be a more difficult task. If such information is not available and it is inferred, the building-level accuracy of these features may be low. Within a given inventory data source, different building features may have different levels of accuracy, as some building features are better documented or easier to capture than others.

In this paper, the *fidelity* of a building inventory refers to the extent to which the inventory data represents the real building stock in terms of accuracy, completeness, spatial resolution, and typology resolution. Creating higher-fidelity inventories involves increasing the spatial and/or typology resolution, while 1) maintaining at least as much accuracy in each building feature as the lower-resolution inventory and 2) ensuring that the data is complete, meaning that buildings are not missing from the inventory. Fidelity represents a balance between the benefits of higher resolution and the potential loss of accuracy for the building features being described. 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 the Hazus inventory to the building level 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 data sources.

When embarking on a regional risk assessment, it is important to consider the purpose of the assessment and whether the additional effort associated with developing high-resolution inventories is justified. Coarse spatial aggregation can bias results, and increasing resolution can help mitigate some of this bias, though this is not guaranteed [61, 55]. Building on this, Babič et al. (2023) shows that the likely benefits of higher resolution can be assessed in advance based on factors such as building stock homogeneity and hazard variation [61]. While there is a benefit, studies focused on aggregate loss metrics indicate that beyond a certain resolution, there are diminishing returns [62]. Because of this, a building-level inventory may be unnecessary for studies focused solely on aggregate losses. In contrast, if the goal of the study is to quantify risk to specific seismically vulnerable structures or critical buildings, identify hot spots, or capture more nuanced risk patterns, high-resolution inventories may be necessary. In these cases, specific vulnerability models can be applied more accurately if the inventory includes sufficient building features to guide model selection, and building-level resolution can capture realistic correlations in building features, which are often 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 accuracy

and completeness to obtain reliable results from the 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 complete and accurate.

2.2. Common Types of Inventory Data

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 data sources that are relevant for inventory development. This section provides a brief overview of commonly used data sources and calls attention to widely-recognized caveats about various public data sources. More specific commentary on individual data sources is included in Appendix C. The discussion primarily focuses on data sources within the United States, but similar concepts can be applied to regions outside the United States. 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 introduction, there are several building inventory data sources 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, Global Earthquake Model (GEM), or Prompt Assessment of Global Earthquakes for Response (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 buildings across the United States [10]. It has been used in regional risk assessments [16, 52] and partially informs the Hazus inventory [63]. 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 data for critical infrastructure across the United States [64]. In general, these national data sources are easy to adopt due to their standardized formats, broad geographic coverage, and relative completeness (typically few to no gaps or holes in the data). However, one challenge in such data is that information likely comes from different sources, making it difficult or impossible to distinguish between what is true data and what has been estimated, potentially limiting local specificity and building-level accuracy.

At the local level, more granular data sources 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 data sources are not designed for natural hazard risk assessment and it takes significant preprocessing effort to convert them into a usable format. Nevertheless, local data sources 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 data sources with global coverage of building footprints, which can be created using computer vision algorithms (e.g., Microsoft Building Footprints [65]), can be community-developed (e.g., OpenStreetMap [66]), or developed by synthesizing multiple individual sources (e.g., Overture Buildings [67]). Most global building footprint data sources 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 data sources 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 [68] currently limits access to aggregated data at the zip code or neighborhood level, which restricts its usefulness for building-specific analysis. While these data sources 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 data sources 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 [69, 70], street-level imagery [71, 72], 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 data sources 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.

2.3. Data Source Classification

To support building inventory data synthesis, a systematic classification is needed to describe how different inventory data sources represent buildings. In this study, we adopt terminology that separates each data source into two components: its source geometry and its building features. First, we refer to the spatial representation of each data source as its *source geometry*, which describes both 1) the geometry (i.e., point vs polygon) and 2) the scope of what the data represents (i.e., single building, multiple buildings, single unit, or a mix of the three). The following list of general source geometry types is used to categorize data sources to inform what they represent and how they can be synthesized. Examples of possible data sources with each source geometry type are provided, but these examples are not exhaustive.

- *Single-building polygons*: a polygon represents a single building (e.g., a building footprint)
- *Single-building points*: a point represents a single building (e.g., a point located at the building centroid)
- *Multi-building polygons*: a polygon represents multiple buildings (e.g., a census block polygon)
- *Multi-building points*: a point represents multiple buildings (e.g., a single point represents an entire mobile home park or university campus)
- *Single-unit polygons*: a polygon represents a single unit (e.g., a tax parcel denoting an individual condominium unit within a larger building)
- *Single-unit points*: a point represents a single unit (e.g., a point located at an individual unit address)
- *Mixed-type polygons*: multiple types of polygon source geometries all listed together within a single data source, such as tax parcels. Tax parcels, which denote ownership, 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*: multiple types of point source geometries all listed together within a single data source, similar to mixed-type polygons
- *Non-spatial data*: Data is not geospatially located and therefore does not have a source geometry

Second, *building features* refer to all characteristics of a building described by a data source, other than its source geometry. Building features include structural attributes (e.g., structural system, number of stories), use-related information (e.g., occupancy type), building value (e.g., replacement cost), 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). The separation of source geometry and building features is also consistent with conventions in catastrophe (CAT) risk modeling, where the spatial location and value of an asset are treated separately from the set of attributes used to characterize its vulnerability [73].

Beyond the source geometry and building features included in a specific data source, it is equally important to consider the data provenance (origin and history). In this study, we adopt three categories to describe data provenance. *Collected inventory data* refers to building inventory data that has been directly observed or recorded through administrative processes, surveys, or measurements, such as property tax assessor records or building permit data. These data sources are often considered closer to ground truth, but they may be incomplete, inconsistently defined across jurisdictions, or missing attributes relevant for hazard analysis. *Estimated inventory data*, in contrast, refers to building inventory data that has been modeled, inferred, or estimated from other information using rules, statistical models, or remote sensing techniques. Such approaches are commonly used to create spatially complete data at large geographic scales, but often have lower inventory accuracy than is typical of collected data. In practice, many building inventory data sources (particularly those with complete information at a broad spatial scale) contain a combination of collected and estimated data, which we will refer to as *uncertain-provenance inventory data*. The challenge with such datasets is not the mixture of collected and estimated data itself, but rather the lack of clarity about which buildings and/or building features fall into each category. This uncertainty makes it difficult to understand the accuracy of each building feature described by the data source.

Recognizing this distinction in data provenance is important when synthesizing multiple data sources, as it affects the likely accuracy of individual building features and how conflicts between sources should be interpreted. It is also worth noting that considerable variation in accuracy can exist within each category (e.g., detailed modeling vs using a homogeneous assumption would both fall under estimated inventory data). Furthermore, there is not a direct relationship between data provenance (estimated, collected, or uncertain-provenance) and accuracy and/or completeness. For example, it is possible to have high accuracy in estimated data, very biased uncertain-provenance data, or many gaps in collected data. These categories are therefore best understood as conceptual classifications to help with interpretation, rather than direct indicators of data quality.

3. Inventory Synthesis Framework

The focus of this study is on synthesizing multiple data sources to create building inventories for use in regional natural hazard risk assessments. The framework presented in this section is intentionally general: it is not tied to a specific hazard, geographic location, or dataset. Instead, it outlines a structured process for combining heterogeneous data sources into a single building inventory that can support a wide range of regional risk assessment applications.

Several decisions regarding the purpose and resolution of the building inventory should guide the selection of input data sources. These decisions influence what building features must be represented in the inventory and which data sources are most appropriate to include. Accordingly, this section first outlines the key inventory design decisions that guide input data selection, then discusses common challenges encountered when synthesizing multiple data sources, and finally presents the proposed inventory synthesis framework for addressing these challenges.

3.1. Inventory Design Decisions

Before synthesizing multiple data sources, several decisions about the intended content and resolution of the building inventory should be made. These decisions help determine which input data sources are necessary and appropriate to include in the synthesis process. In practice, these choices should be guided by the purpose of the overall regional risk assessment. The key considerations are outlined below:

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 coarsest (simplest) spatial and typology resolution necessary to address study 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 vulnerability model, evaluation of impact). It is critical to evaluate each data source for location availability (whether it covers the area of interest), completeness (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 the origin of the data, including the creator, host, purpose, creation date, last update, and the point in time that the data most accurately reflects. The database's original purpose likely determines what is prioritized during its development as well as its provenance (estimated, collected, or uncertain-provenance data). For example, national data sources like NSI are designed to be complete and contain very few gaps, but achieve this by filling missing data through disaggregation and inference, resulting in largely uncertain-provenance data. Conversely, local tax databases are an example of collected data sources, which 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:** To synthesize multiple input data sources into a single inventory, all input data must be co-located into a common *baseline geometry*, which is the selected spatial representation used to describe the final inventory. The selected baseline geometry should be appropriate to the study's spatial resolution. For example, if the targeted spatial resolution is building-level analysis, the selected baseline geometry should be at the single-building level. This study recommends using single-building polygons as the baseline geometry for several reasons. First, polygons enable easier linking between data sources, especially for large buildings where point data may be scattered; linking is more difficult when the selected baseline geometry is point-based. Second, single-building polygons have physical meaning; they represent individual buildings with consistent building features. This is not true of all polygon source geometries; for example, tax parcel polygons, which denote ownership, may correspond to a single building, a single unit within a larger

building, or multiple buildings (i.e., they have mixed-type polygon source geometry). Using parcel polygons as the baseline geometry can make it challenging to track building features, such as when a single parcel encompasses multiple buildings with different structural systems. While errors may occur in delineating single-building polygons (e.g., adjacent buildings incorrectly merged), these reflect data quality issues that can be screened for, rather than inherent limitations of the representation. Third, specifying the inventory at the individual building level preserves correlations between building features that would otherwise be lost through aggregation (e.g., if a multi-building polygon such as a census block was used as the baseline geometry). When selecting a baseline geometry source, it is important to screen for its quality by comparing it to other possible sources, as any gaps or errors in the chosen source will produce corresponding gaps and errors in the final inventory.

Once the above inventory design decisions have been made, the remaining task is to synthesize the selected data sources into a building inventory. This paper introduces an inventory synthesis framework as a standardized approach, and the remainder of this section introduces the associated challenges and steps in this process.

3.2. Data Synthesis Challenges

An important and well-known challenge in synthesizing multiple data sources is that different sources represent buildings differently both in space (i.e., they have different source geometries) and in content (i.e., they have different building features). Thus, the process of co-locating data into a common 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 two example data sources, including NSI point data and a building footprint data source. Figure 1a shows NSI points that fall just outside of the corresponding footprint (which is problematic for spatial joins based on overlap and intersection), cases where multiple points fall within a single footprint (which can arise if points denote individual units within a single building), and cases where footprints have no associated points. Figure 1b shows an example where points do not clearly relate to any footprint, including cases where they fall along transit lines (see expanded view in Figure 1c). Similar issues can also arise between data sources with different polygon source geometries, such as building footprints and tax 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 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, resulting in data being dropped from the inventory or being inconsistently handled. Carefully attributing data to a common baseline geometry reduces the risk of data being inadvertently dropped and can help produce more robust results.

Beyond the spatial challenges of co-locating multiple input sources to a common baseline geometry, additional challenges arise from differences in the available building features and the way they are described. There may also be disagreement between sources, gaps in the data, and other issues that must be resolved through the process of data synthesis.

3.3. Inventory Synthesis Framework

In light of these challenges, this paper introduces the inventory synthesis framework shown in Figure 2, which provides a generalized approach for synthesizing heterogeneous data sources into a consistent building inventory. Five main steps are identified, including preprocessing data, attributing input data sources to the selected baseline geometry, selecting values for features with disagreement, filling gaps, and mapping to features required for simulation. These steps are not tied to a particular hazard, location, or dataset and can be adapted to different contexts.

The framework was originally developed using single-building polygons as the baseline geometry. In general, the most widely-available single-building polygons are building footprints. Thus, it is assumed throughout the remainder of the study that footprints are used as the baseline geometry. If an alternate baseline geometry is selected (e.g., multi-building polygons), some modifications to the framework may be required.

The first part of the inventory synthesis framework is to preprocess the data, which involves four main steps (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

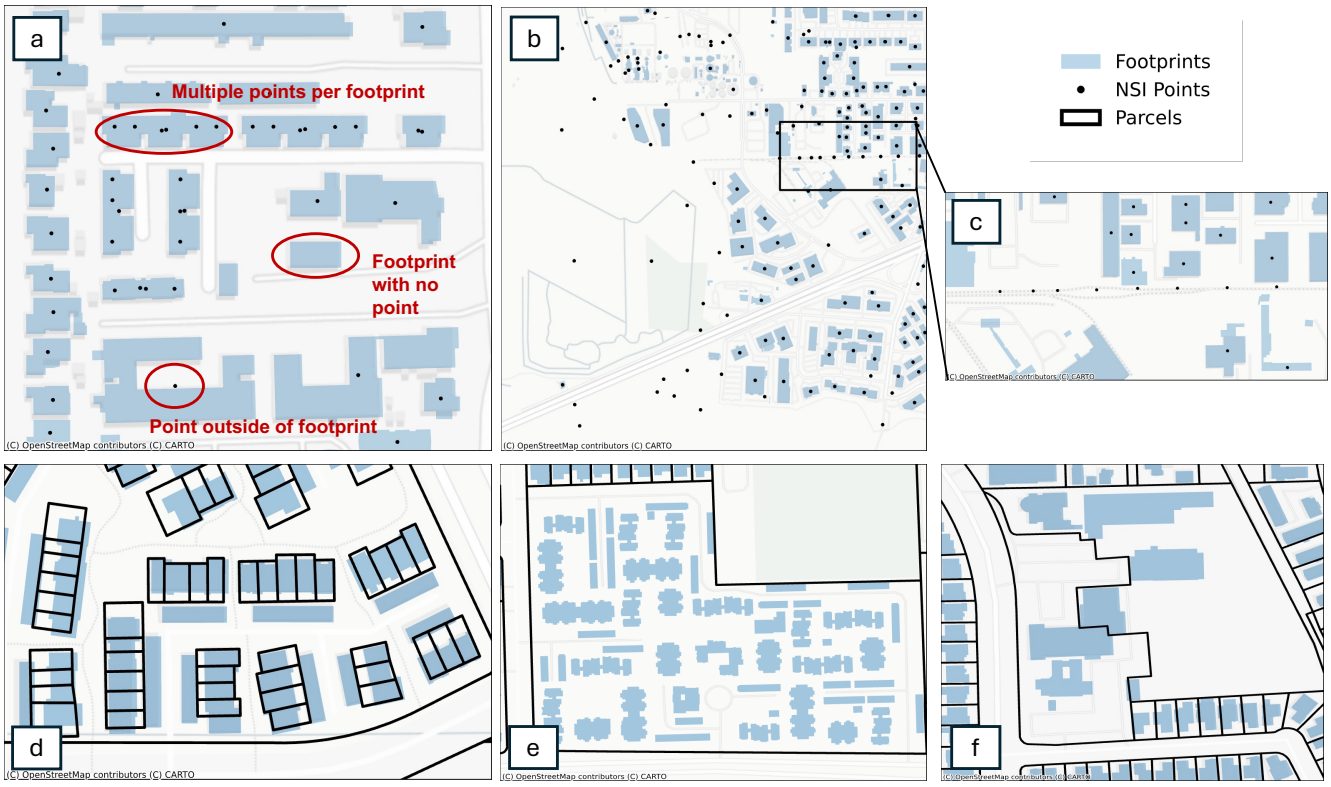

Figure 1: Examples of approximate geometries and lack of one-to-one mapping across source geometries for points and footprints (a-c) and parcels and footprints (d-f). The point data shown here is from NSI [10] and the footprint and parcel polygons are from the case study discussed in Section 4 (local data from Hayward, CA).

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. Third, 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.

Following preprocessing, the next major step is to attribute all input data sources to the selected baseline geometry, building footprints (Figure 2, Box B). In a specific implementation of the above framework, input data should be classified based on source geometry, as different source geometries require different methods for attributing data to footprints. Prior to the attribution process, one input should be identified as the *bounding geometry*, which is a polygon geometry used to assist in attributing data to footprints. The bounding geometry polygons should be non-overlapping and have complete coverage across the study area (e.g., census blocks or tax parcels). While there are computationally efficient tools for assessing direct intersection (overlap) of footprints with other point and/or polygon data, the attribution methods described in this study aim to accommodate the approximate geometries and lack of one-to-one mapping shown in Figure 1. Accommodating these issues requires computing distances between different geometries. In a city with thousands or hundreds of thousands of buildings, this can be computationally demanding. The purpose of the bounding geometry is to improve efficiency and provide reasonable distance limits for attributing data to building footprints.

Thus, the process of attributing data to footprints is completed in two steps. First, all input data, including the baseline geometry footprints, is tagged with the bounding geometry polygon it falls within. For example, if census block polygons are being used as the bounding geometry, all input data would be tagged with the appropriate census block number. Second, all input data is attributed to footprints. By first assigning data to a corresponding bounding geometry (e.g., census block), distance calculations can be limited to footprints within the same (or adjacent) census blocks, thus improving efficiency in the method. The use of adjacent census blocks accommodates edge cases where a footprint spans census block boundaries. In addition, the bounding geometry can also provide

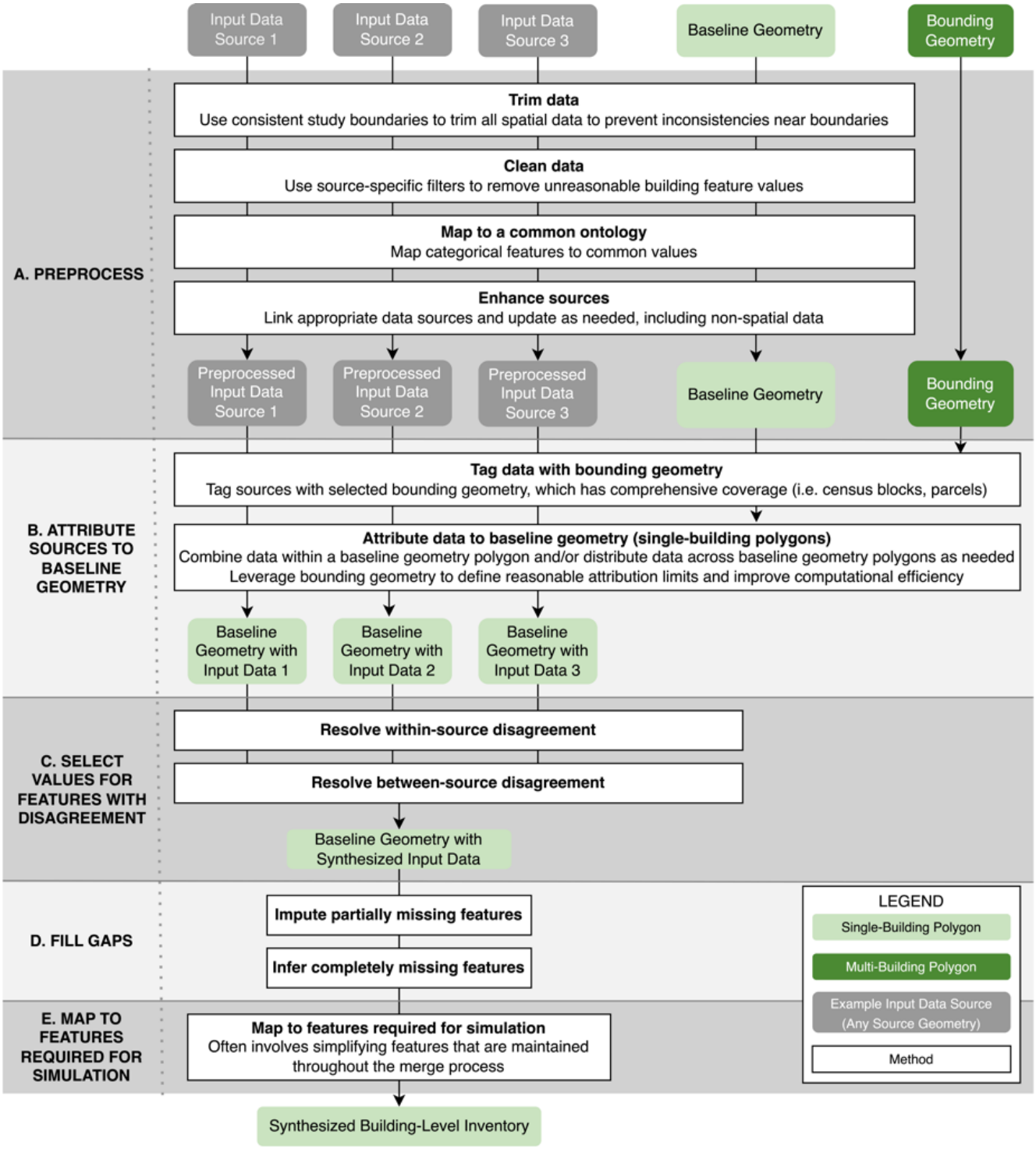

Figure 2: Inventory Synthesis Framework

reasonable bounds for how far data can be moved to be attributed to a footprint (e.g., data cannot be attributed across a census block boundary line). More details on how bounding geometries are used are described in Table 3. The result of this process is that all input data is attributed to a consistent set of building footprints (i.e., the baseline geometry of the inventory), which allows for information to be compared and synthesized across multiple input data sources.

In a synthesized building inventory, each footprint should have a single value for each building feature. For example, each footprint should have a single value assigned as the number of stories, a single value assigned as the

occupancy class, etc. However, when multiple data sources are synthesized into a single inventory, it is common to find the same building feature in more than one data source. In this case, the corresponding feature values often disagree within each footprint. Thus, the next step of the framework is to select values for features with disagreement (Figure 2, Box C).

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 different sources report conflicting information about a given building feature of a single building footprint. This occurs when one or more sources contain inaccurate 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 between-source disagreement arises when one or more data sources derive building-level values by disaggregating block-level or tract-level statistics. In such cases, data may agree at an aggregated spatial scale (e.g., a census tract) but disagree at the footprint level. For example, a data source may assign a structural system to each building by sampling from census-tract-level proportions. While the tract-level proportions of different structural systems may be correct, the assigned structural system for a specific building may disagree with another data source that provides building-level information. Between-source disagreements can be resolved by prioritizing one source over another.

In some cases, within- or between-source disagreements are not conflicts, but instead demonstrate the need for a more sophisticated building ontology. 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 ontology of the input data sources is not sufficiently granular for describing that complex occupancy class. 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 ontology. In either case, it is important to acknowledge the uncertainty revealed by the disagreement in the data.

The next step of the synthesis framework is to fill any gaps in the required building features (Figure 2, Box D). We recognize two types of gaps: partially missing and completely missing.

A feature is *partially missing* when it is available for a subset of the buildings. Partially missing information can be either random or systematic. Random gaps occur without a predictable pattern in otherwise available features and can be addressed through *imputation*, where missing values are estimated using patterns observed in the existing values of that same building feature [18, 31, 74]. However, in realistic inventories, data is often missing systematically, not randomly. Systematic gaps occur in a predictable pattern; for example, all public educational facilities may be missing the year built due to differences in data collection such as tax-exempt status. Filling systematic 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 that feature were completely missing.

A feature is *completely missing* when it is not available for any building in the inventory. In this case, it cannot be imputed because there are no available observed values of that feature. Thus, completely missing data must be *inferred* using external models or reasonable assumptions based on other available features. For example, the structural system of a building can be inferred using information on its height, occupancy type, and year built.

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 from 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 at the end of the framework, more detailed data can be tracked throughout the process and potentially used for other types of analyses.

The framework presented here aims to create deterministic inventories by synthesizing multiple data sources. However, it is also important to acknowledge uncertainty in the building inventory. Unlike some sources of uncertainty in regional risk assessments (e.g., ground motion record-to-record variability), inventory uncertainty is epistemic, stemming from insufficient data about the physical world, not inherent randomness. The proposed framework currently resolves disagreements and fills gaps through the assignment of a single building feature value, masking uncertainty. However, the proposed framework could be extended to quantify and propagate inventory uncertainty, and ongoing work by the authors is exploring this topic.

4. Methods and Illustrative Case Study

Based on the inventory synthesis framework in Figure 2, this section proposes specific methods for handling each step. We illustrate all proposed methods with a concrete example using data from Hayward, California, to provide a clearer explanation. Three different workflows based on the framework in Figure 2 are presented: 1) the national synthesis workflow, which uses only nationally available data, 2) the local synthesis workflow, which uses only locally obtained data, and 3) the best estimate workflow, which combines 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 adapting scripts to the schema of local tax records and other location-specific data sources. Prior to synthesizing data, this section considers the inventory design decisions 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 the Hazus vulnerability modeling framework, which uses fragility functions to estimate damage as a function of earthquake intensity. Thus, this study requires a building-level spatial resolution and a typology resolution sufficient to inform Hazus vulnerability model selection. Building features required for model selection include the structural system, height class (mapped from the number of stories), and seismic design level (mapped from the year built). Structural system is often not available in common data sources, so Hazus provides building stock mapping scheme tables [75], which use the region, occupancy class, year built, and number of stories to assign probabilities for different structural systems, which can be sampled. Building material, if available, can be used to further constrain the possible structural system. Beyond vulnerability model selection, the number of residential units, population, and replacement cost can be useful for assessing risk and impact. Thus, based on the selected spatial and typology resolution, the targeted inventory requires building-level information on occupancy class, year built, number of stories, structural system, building material, number of units, population, and replacement cost.

Select data sources: The goal of input data source selection is to assemble sources that collectively contain all the required building features. The Hayward example study draws from both national and local data sources to do so. At the national level, the primary data source is NSI. HIFLD is also incorporated to complement NSI and provide higher-quality location information for police and fire stations, emergency response centers, schools, colleges/universities, and mobile homes. In this 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) [76]. The California School Directory is used to supplement school features from HIFLD with building year built data.

Data sources are summarized in Table 1 and more descriptive information is provided in Appendix C. As shown, all required features are available from at least one data source apart from structural system, which is not available at all. This is common, as information on the structural system of a given building is not typically available in public data sources. Without it, structural system must be inferred using other features. This presents a major limitation in available data because structural system is a primary driver of seismic vulnerability and loss. In addition, building value, in the form of 'improvement value,' was available in the parcel data, but due to its variability based on year of purchase and other factors, it was not included as an input in the data synthesis.

Select baseline geometry: Consistent with the Section 3, this step adopts 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 understand 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 complete national data source but it was missing recently constructed buildings, visible in Figure 3, View C. Overture included newer construction and had more accurate

Table 1: Inventory data sources used in the Hayward case study, including their source geometry and available building features.

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

Note: An asterisk (*) indicates the feature is directly available. A description indicates one of the following cases: 1) 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) a different feature can be used as a proxy for approximation (e.g., height).

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 complete and was selected for use in 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 coverage of the Hayward footprints, this was not included in the scope of this study.

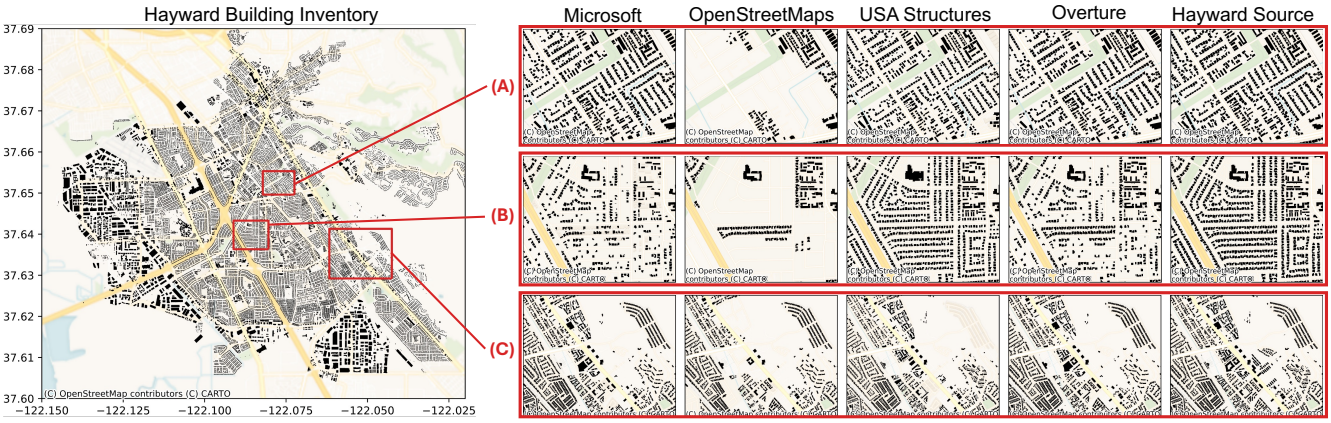

Figure 3: Coverage of five different footprint sources for different locations in Hayward.

Once these decisions have been made, the remaining task 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. National Synthesis Workflow

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 for Hayward using only nationally available data sources. Input data includes NSI, HIFLD, census block data, and Hayward building footprints (see Table 1 for details). The highlighted methods in Figure 4 denote more detailed methods proposed in this study that are outlined explicitly in the text.

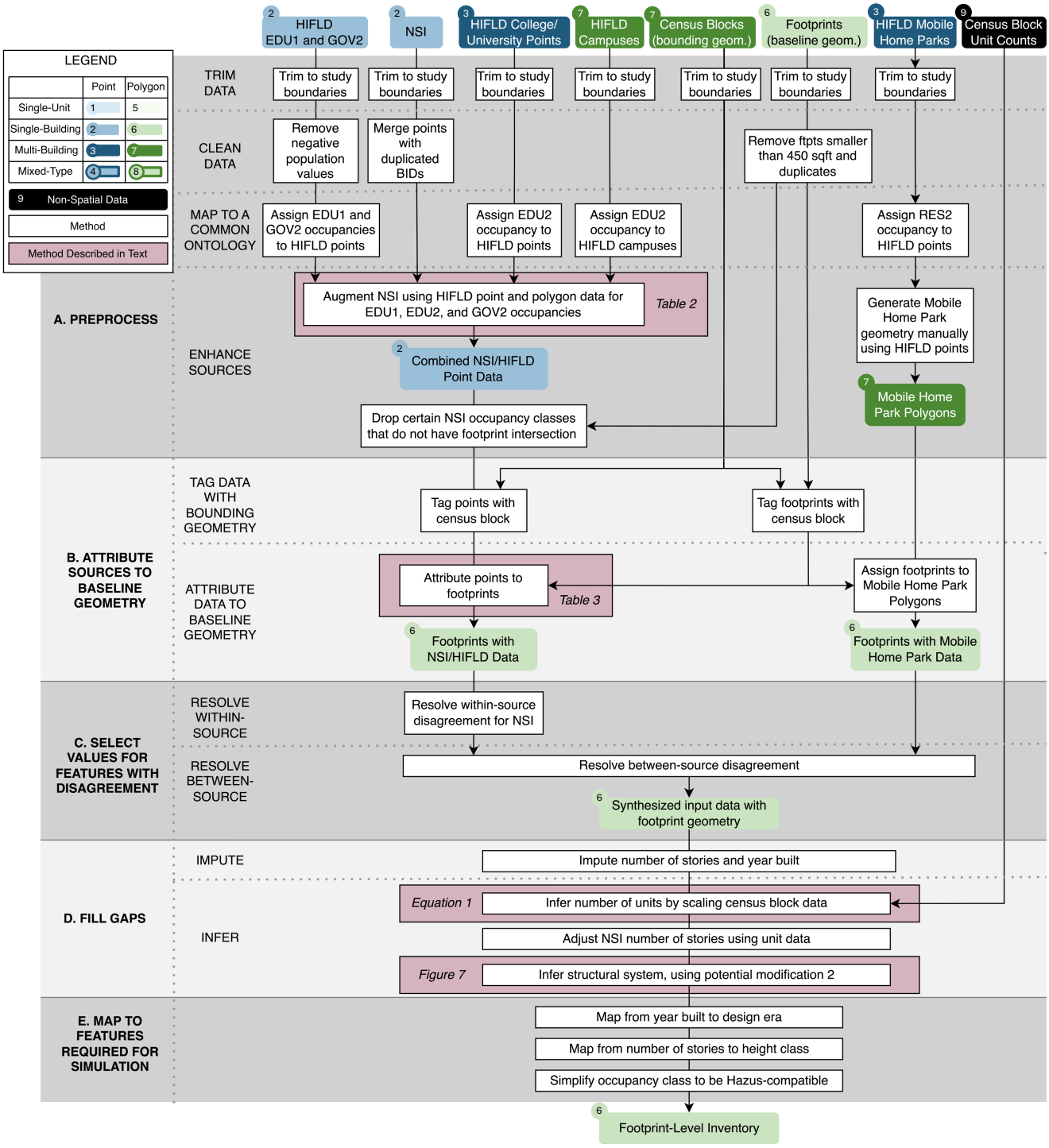

Figure 4: National Synthesis Workflow

Preprocess Data (Figure 4, Box A)

The first step in preprocessing is trimming data to consistent study boundaries. Here, 2010 United States 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, the same size threshold applied in the USA Structures dataset, as these

typically correspond to sheds or other small buildings, and (2) removing overlapping and duplicated 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, an 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 material, 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 B.

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 preprocessing step aims to 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 a polygon 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 address this by removing these points.

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 was 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 data must be attributed to the selected baseline geometry of the inventory (footprints). For NSI and HIFLD, 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. This method is relevant across cases where single-building or single-unit points must be attributed to building footprints. Standard methods such as spatial intersection joins often drop valuable data because the points do not explicitly overlap with a footprint, and 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. In the case of the national synthesis 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

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 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 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.

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 classes 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.

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

Table 3: Point-to-footprint attribution method

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 (e.g., dollar value, population) and 2) features that cannot or should not be summed (e.g., 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 (e.g., 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. These adjacent footprints 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 may not be 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 [63]. 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-unit data, where each point represents one unit, but occupancy is often tagged for the entire building (e.g., one unit in an 8-unit building may 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 5, 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 the national synthesis workflow, footprints considered in this step are limited to the bounding geometry of interest (not adjacent) to avoid attributing potentially erroneous points such as those shown in Figure 6.

Step 7: Drop remaining points
Remove any remaining unassigned points from the inventory.

to ensure that buildings that may have higher impact, such as a 50+ unit residential building, are not overwritten by another possible option in the footprint. If a residential occupancy is paired with a non-residential occupancy,

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

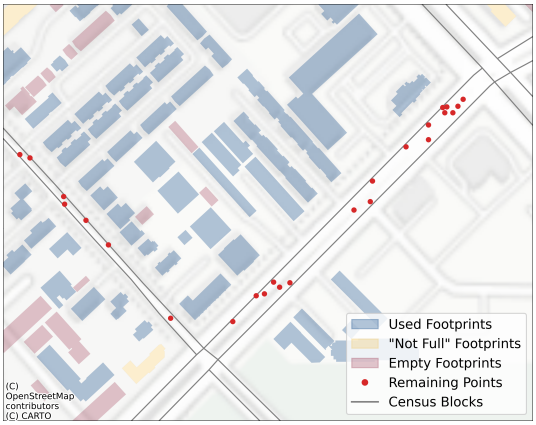

Figure 6: NSI points located along road-only census blocks, dropped at the end of the workflow.

an "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 a more reliable source of occupancy class information based on spot-checking points in Hayward, it is prioritized over NSI when occupancy class disagrees. Similar procedures are applied for other building features as well.

Fill Gaps (Figure 4, Box D)

Due to the completeness of the NSI data, there are no features that are partially missing at random. However, HIFLD data is systematically missing year built, number of stories, and building material; that information is simply not included in the HIFLD data. 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 existing 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 vulnerability model 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 explores 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 scaling factor shown in Equation 1 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, no data source reports the structural system, which is a required building feature for selecting a vulnerability model and assessing seismic loss. Thus, this feature is completely missing and must be inferred. Structural system is inferred using the Hazus building stock mapping scheme tables [75], which provide region-specific percentages of floor area for specific structural systems, given other available building features [9]. These percentage values can be used to sample structural system probabilistically, and a schematic of this sampling process is shown in Figure 7.

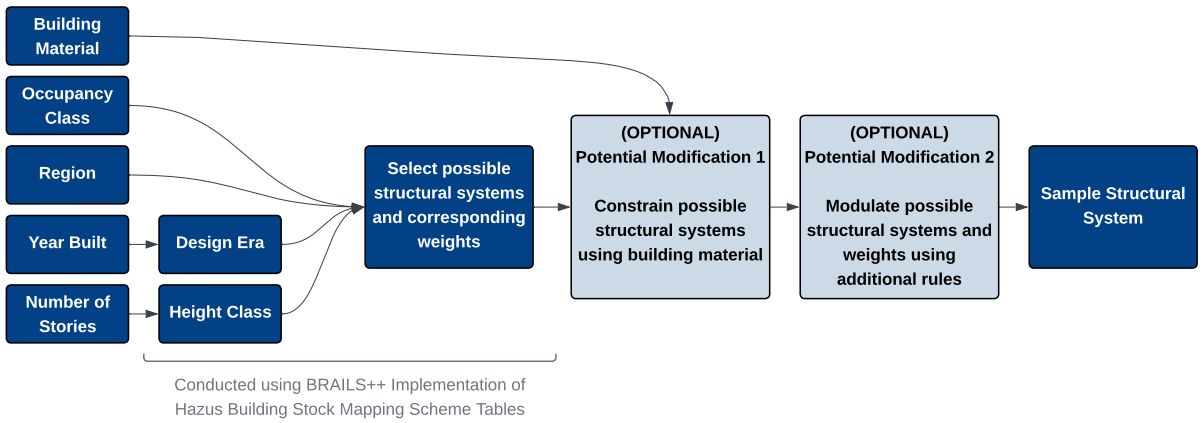

Figure 7: Schematic of inferring structural system based on other available building features. Based on the Hazus Building Stock Mapping Scheme Tables with two potential modifications [75].

As shown in Figure 7, the first step uses built-in tools in BRAILS++ to select possible structural systems and corresponding weights based on occupancy class, region, year built, and number of stories according to the Hazus building stock mapping scheme tables [18, 75]. Based on those structural systems and weights, it is possible to sample the structural system directly. In this workflow, we consider two possible modifications.

First, it is possible to constrain the structural systems using building material. For example, a low-rise medical office building on the West Coast built between 1950 and 1970 could have Hazus structural system W2, S1L, S2, S3, S4, C2L, RM1L, or URML, meaning that the building could have a variety of structural systems ranging across various steel, concrete, wood, reinforced masonry, and unreinforced masonry systems (see Hazus documentation for interpretation). These values each have a corresponding weight, which can be used to sample probabilistically. If the building material is known to be steel, the possible structural systems can be reduced to S1L, S2, S3, S4, which can then be sampled according to their weights. Although building material is available in the NSI data, 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, structural system assignment was done without regard to the NSI building material, meaning that the first potential modification shown in Figure 7 is not adopted in the national workflow. For example, even if an NSI residential building is listed as having masonry as the building material, possible structural systems are not limited to only masonry systems.

The second optional modification is to modulate the possible structural systems using three additional rules, which were found to be beneficial in the context of Hayward. First, structural systems 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 buildings. For example, it is unlikely for a 2-unit duplex to be a steel structure, which could otherwise occur if all multi-family (RES3) buildings are treated the same, regardless of the number of units. Second, the MH structural system in Hazus was only assigned for mobile homes (RES2), rather than being a structural system option for multi-family (RES3) occupancies. Finally, no unreinforced masonry (URM) structural systems were assigned in Hayward because of the 1986 URM retrofit program, which requires local governments in high seismic zones to identify and establish loss reduction programs for URM buildings [77]. Thus, the second potential modification shown in Figure 7 is used for the national workflow.

Even with standardized methods to infer structural system, a significant amount of uncertainty remains after this step due to the variability in structural system across occupancy classes and the absence of standard methods to collect and report this information. In the case of the national workflow, sampling structural system based on the Hazus mapping scheme tables is an example of a method that increases the spatial resolution of the resulting inventory without substantially increasing the fidelity of it, due to the low building-level accuracy of structural system. This is further discussed in Section 5.

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

The inventory in this study targets a regional simulation using Hazus vulnerability models. 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 classes, which allow 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 national synthesis workflow and results in a building inventory that can be used for downstream regional risk assessment studies.

4.2. Local Synthesis Workflow

The following discussion demonstrates how the general framework in Figure 2 is applied to create a footprint-level inventory using local data sources. The local workflow for Hayward is shown in Figure 8. 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 8, Box A)

The same preprocessing steps used in the national workflow – 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 data entry errors or inconsistencies 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 to the ontology. 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

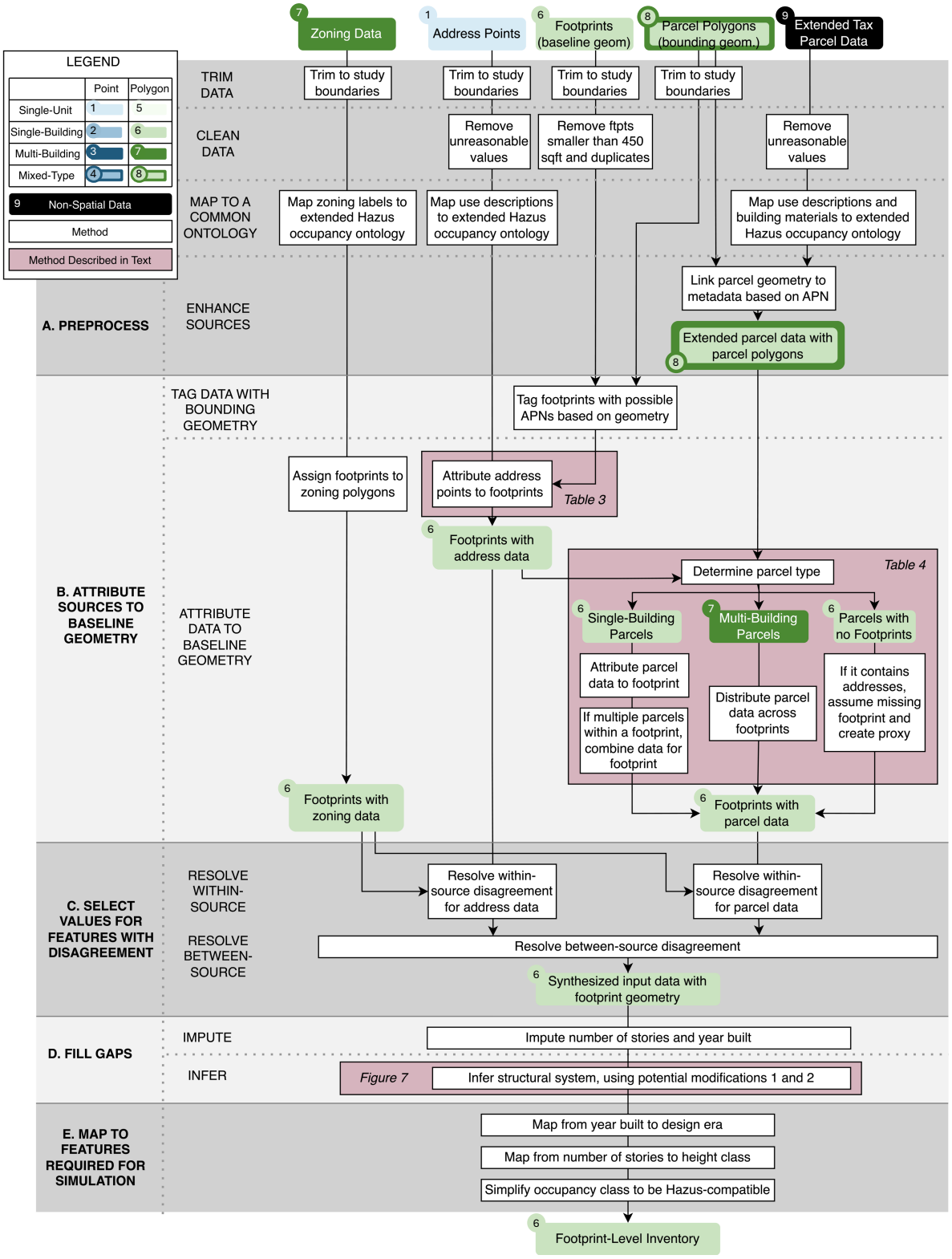

Figure 8: Local Synthesis Workflow

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 B.

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 data sources.

Attribute Sources to Baseline Geometry (Figure 8, 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 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, as such overlap typically only results from slight misalignment in the geometry. 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 9.

Table 4: Parcel-to-footprint attribution method

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 9c), store parcel data either by summing (dollar value) across parcels associated with the same footprint, or by storing single values or lists of values for features that should not be summed (e.g., number of stories, building type).	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 (e.g., occupancy class), or by dividing data across all footprints in the parcel (e.g., 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 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 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 9 shows an example of the cases described in Table 4. Case 1, where a parcel corresponds to one footprint with address data, can include single-family homes (9a), single-family homes with sheds or garages that do not have addresses (9b), or a single unit within a larger building (9c). 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 (9d) to many footprints with many addresses, such as a mobile home park under a single owner (9e). Figure 9f shows an example of Case 3, where a parcel corresponds to one or more footprints, but does not correspond to any address points. Finally,

Figure 9g 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.

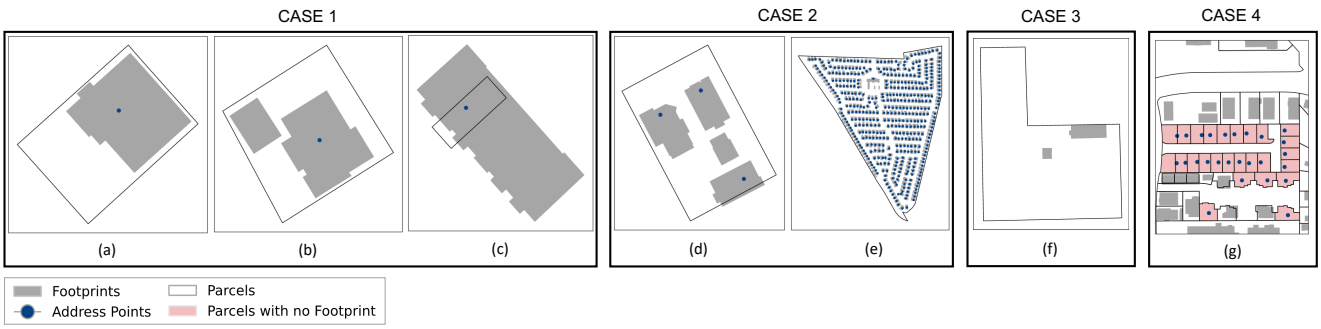

Figure 9: Examples of parcel, address point, and footprint geometries that correspond to each case described in Table 4.

Select Values for Features with Disagreement (Figure 8, 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 complete coverage nationally (NSI), two sources (tax parcels and address points) have 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 8, 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. In addition, 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.

No local data source provides information on structural system, and the closest available proxy from the local data is the building material. Thus, structural system is inferred using Figure 7, as in the national workflow. However, in this case, the building material from the local data is used to constrain the possible structural systems (i.e., both the first and second potential modifications are adopted in the local workflow). While building material does not fully specify structural system, it significantly reduces the range of possible structural systems for each building, likely improving building-level accuracy of this assignment. This increase in accuracy allows for a corresponding increase in the fidelity of the resulting inventory, though fidelity remains limited compared to what could be achieved with collected building-level structural system data.

Map to Features Required for Simulation (Figure 8, Box E)

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

4.3. Best Estimate Workflow: 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, which is 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 10. There were 849 such footprints across Hayward, and the NSI points attributed to these footprints represented 416 million dollars and 10,370 people. 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 removing these footprints from the baseline geometry. A similar adjustment is made to the baseline geometry to incorporate proxy geometries (5-meter circles) for missing footprints like those in Figure 9g.

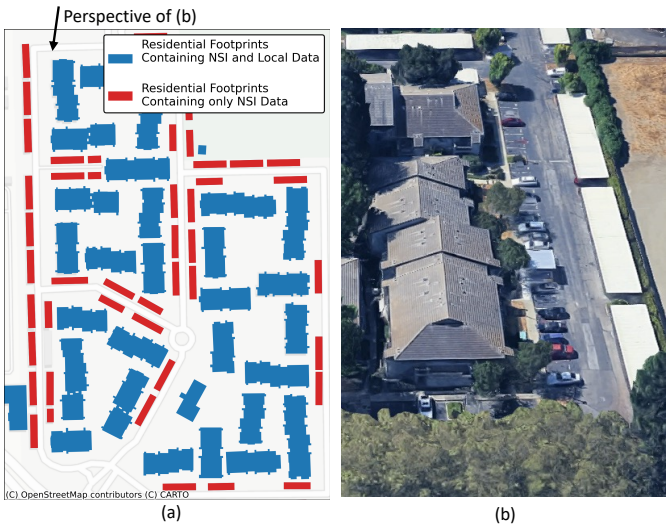

Figure 10: (a) Residential footprints containing both NSI and local data (blue) or containing only NSI data (red). Based on the Google Maps satellite image shown in (b), footprints containing only NSI data seem to be covered parking structures, not actual residential buildings.

The best estimate workflow can be visualized using Figures 4 and 8. First, Steps A and B of both the national and local workflows are run to preprocess data and attribute all sources to the selected baseline geometry (footprints). Then, results are compared and used to update the footprint source, as described above. The updated footprints are then used as an input to the national workflow to re-attribute national data (Steps A and B). This allows for the value and population to be redistributed to nearby footprints, without dropping data or incorrectly attributing data to unrealistic structures. While this may not significantly 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* preprocessing step. 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 values of year built 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 observed 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 collected rather than estimated through national-level assumptions, likely resulting in higher accuracy at the footprint level. Figure 11 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 do not 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 11a shows a school assigned by HIFLD, residential buildings assigned by address points, and smaller buildings around the school, which do not have local data, assigned by NSI. Though address point data is available for most residential structures, Figure 11b highlights greater variability in industrial and commercial areas, where occupancy class is assigned from mixed sources based on availability.

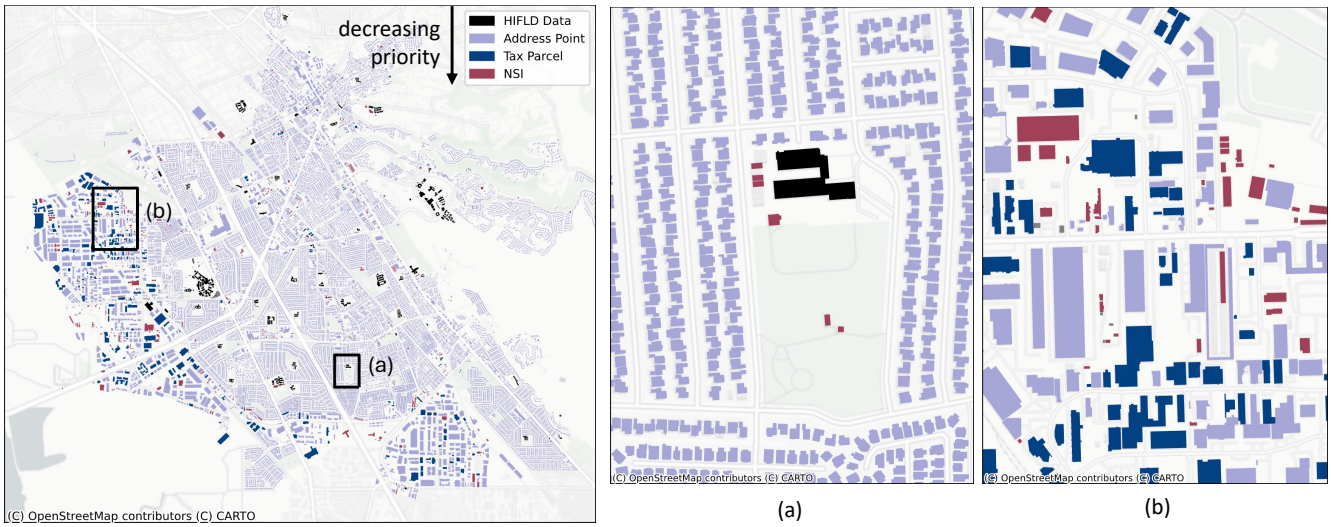

Figure 11: Example of Source Prioritization for Occupancy Class.

Imputing and inferring gaps in the data follows a similar process to the national and local workflows. If the building material is available from local data, it is used to inform which structural systems should be sampled (per Figure 7). If not, all structural systems 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, including height class, design era, and simplified occupancy classes, are used in the best estimate workflow.

5. Implications and Illustrative Results

5.1. Implications for Building Features

Through comparisons of summary statistics and individual building features including occupancy class, number of housing units, and building material, this section demonstrates that the choice of input 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 Section 4 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]. In both the *NSI Point* and *NSI Spatial Join* inventories, structural system is sampled using the method shown in Figure 7, using both potential modifications. Appendix D summarizes the data provenance across the five inventories, indicating whether each building feature is primarily characterized by collected data, estimated data, or uncertain-provenance data.

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 12. 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, and 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 complete 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.

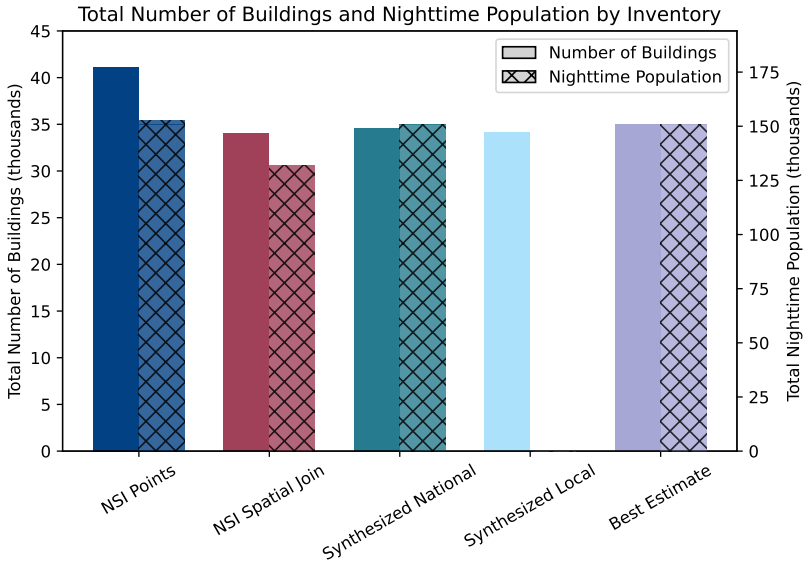

Figure 12: Total number of buildings and total nighttime population by inventory. The *Synthesized Local* inventory does not have a population estimate because there was no local source with population data available.

General patterns in occupancy class are similar across inventories, but there are a few important differences shown in Figure 13a. 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 single-family (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 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 data provides the most complete representation of governmental and educational buildings and is 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 13b illustrates that disagreements are less common when using broad occupancy class categories (e.g., RES, COM), whereas Figure 13c shows more widespread disagreement when comparing specific occupancy classes (e.g., RES1, RES3A, COM2). Figures 13b and 13c 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

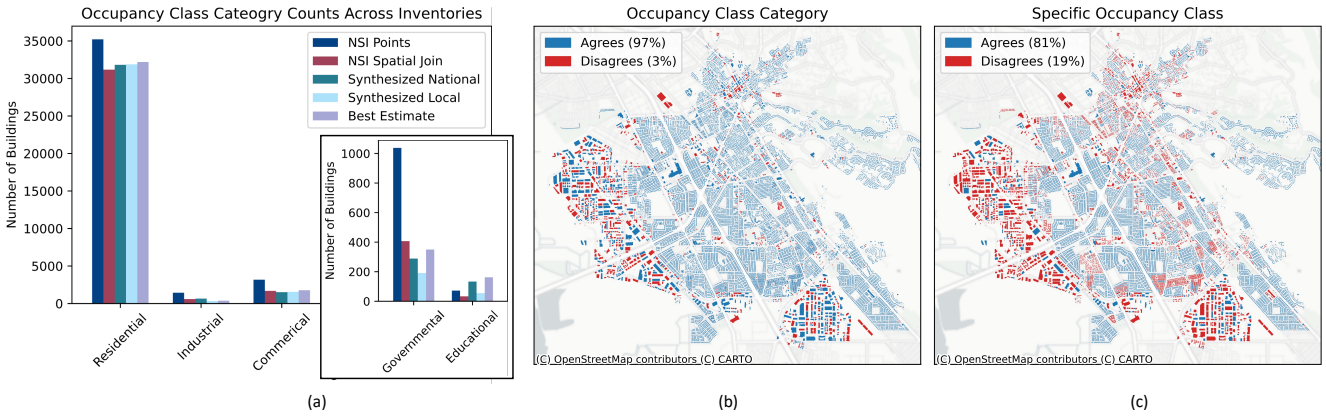

Figure 13: (a) Overall building counts by occupancy class category in each inventory. Footprint-level agreement of (b) general occupancy class category and (c) specific occupancy class between *NSI Spatial Join* and *Synthesized Local* inventories.

(RES1), mobile homes (RES2), and multi-family housing (RES3A-RES3F). Figure 14a 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 far 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). As mentioned, 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 14b and 14c illustrate that disagreements in footprint-level occupancy class values for residential buildings tend to cluster geospatially. Figure 14b compares the *NSI Spatial Join* against the *Synthesized Local* inventory, while Figure 14c 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 data sources can produce a more accurate approximation 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 14d and 14e 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 14d), 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 14e 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 the structural system for these occupancy classes, recognizing that these buildings are often constructed in similar ways. This assumption relaxes the accuracy needed in classification between RES1, RES3A, and RES3B occupancies (single-family, duplex, and 3-4 unit residential buildings, respectively).

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 reason is 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 [63] are also in agreement and underestimate RES2 counts by a factor of two to three. 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.

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 buildings compared to local data. The census unit scaling method introduced earlier and formalized in Equation 1 addresses both issues. Figure 15 shows the correlation coefficient between the number of

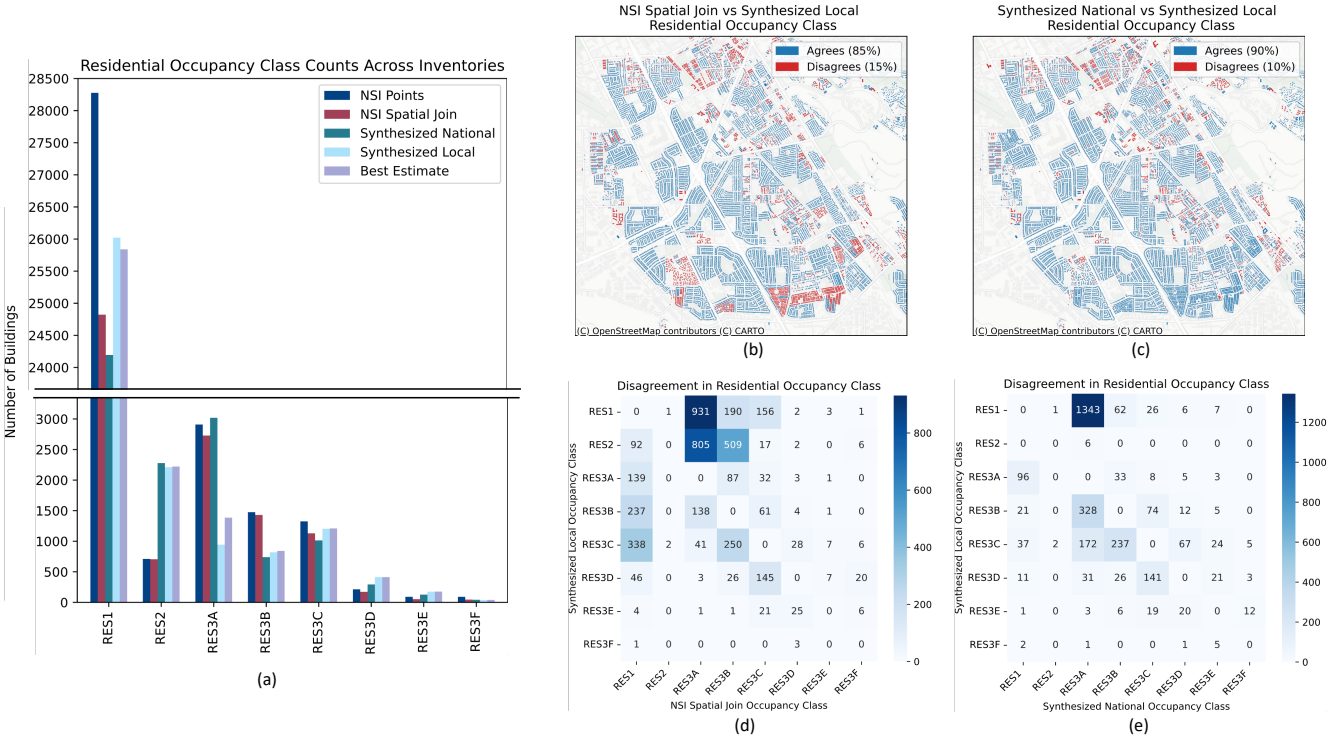

Figure 14: (a) Building counts by residential occupancy class in each inventory. Footprint-level agreement of residential occupancy class comparing the *Synthesized Local* inventory with the (b) *NSI Spatial Join* and (c) *Synthesized National* inventories. Confusion matrices for disagreeing footprints comparing the *Synthesized Local* inventory with the (d) *NSI Spatial Join* and (e) *Synthesized National* inventories.

Table 5: Number of Mobile 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 Point Inventory	244	123	211
Hazus 6.1 Inventory	245	128	211

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 (15a), the average of the NSI unit range (15b), the minimum of the NSI unit range (15c), and the maximum of the NSI unit range (15d). The Equation 1 census unit estimation method (15a) shows the strongest correlation with the *Best Estimate* Inventory across the investigated methods for the *Synthesized National* Inventory. In addition, Figure 15 shows that the total number of RES3F buildings estimated using Equation 1 (15a) is the closest approximation of the number of RES3F buildings according to the local data. The census housing unit scaling method overestimates RES3A buildings compared to the local inventory because RES3A serves as a lower bound for unit assignment in multifamily buildings. This overestimation has minimal impact on the risk assessment since the structural system 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 16a, classifications for building material, including wood (W), mobile homes (H), masonry (M), concrete (C), and steel (S), 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 16b, building material disagreements do not appear geospatially clustered.

In the case of the *Synthesized National* inventory, structural system is sampled according to Figure 7 without

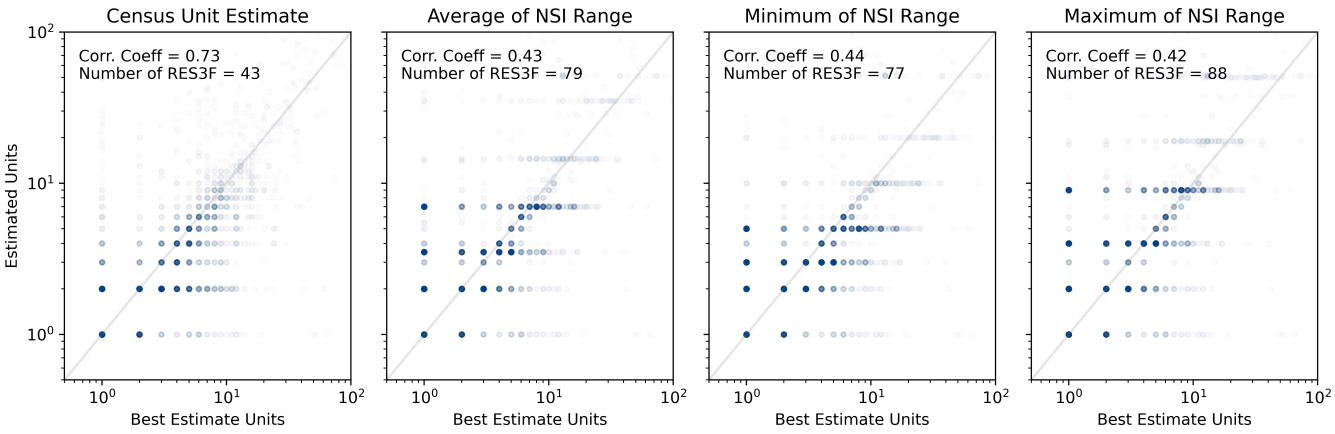

Figure 15: Comparison of housing unit estimation methods from the national workflow, using housing units from the *Best Estimate* inventory as reference. Each plot displays the corresponding correlation coefficient and the number of 50+ unit multi-family (RES3F) buildings. There are 40 RES3F buildings in the *Best Estimate* inventory.

considering the NSI-assigned building material. Based on Figure 16a, building material designation from NSI is biased, greatly overestimating the presence of masonry; thus, including it in the structural system sampling would result in biased selection of Hazus vulnerability models. If structural systems are sampled without regard to NSI building material, building material can be extracted back from the selected structural system (i.e., a footprint with selected structural system C1 would have building material C). For commercial and industrial buildings, sampling structural system without considering NSI building material does not significantly improve building material classification; however, it notably improves residential classifications, correctly identifying most 1-4 unit homes as wood framed. This can be observed in Figure 16c. While the commercial and industrial footprints on the city's west side still see high levels of building material 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 material at the footprint level, which can in turn be used to sample appropriate structural systems. Furthermore, since structural system assignment also depends on the year built, occupancy class, and number of stories, improving the accuracy of those features through the use of local data can further improve the accuracy of the structural system assignment (per Figure 7). While this is not the same as having building-level structural system data, it does reduce some of the uncertainty in sampling structural system.

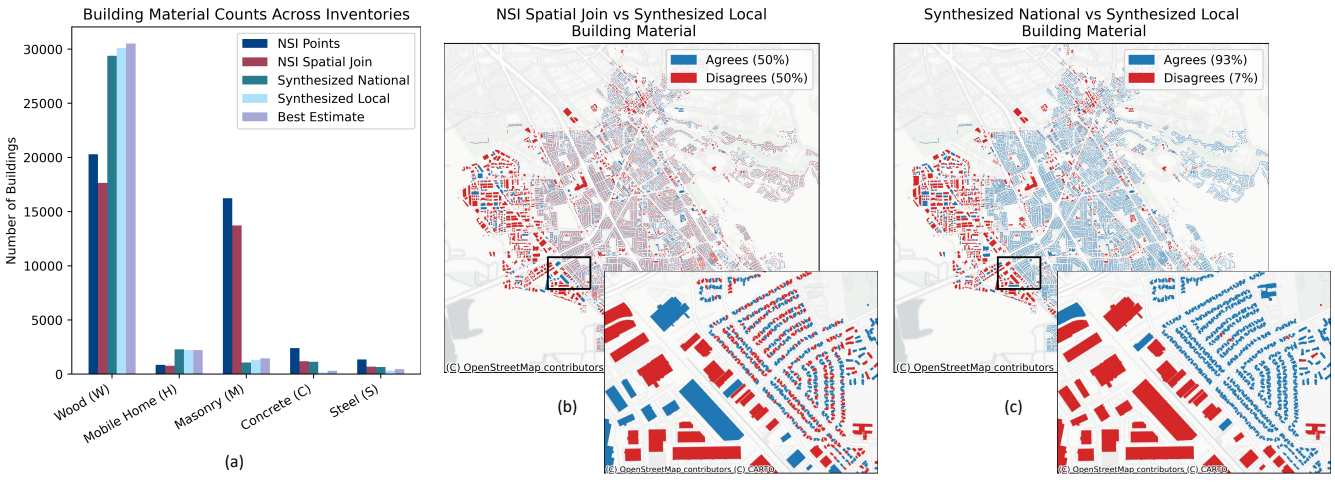

Figure 16: (a) Building material count by inventory. Footprint-level building material agreement between (b) *NSI Spatial Join* and *Synthesized Local* and (c) *Synthesized National* and *Synthesized Local*.

Regardless of sampling technique, structural system assignment for all five of the Hayward inventories remains limited for two main reasons. First and foremost, structural system accuracy is limited because that data is not

available at the building level from any national or local data source. This is often the case and is not unique to Hayward. Second, structural system is being sampled using the Hazus building stock mapping scheme tables, which are specific to a given region (e.g., West Coast) and era (e.g., pre-1950) [75]. This is a large limitation because we are assuming that the structural systems and corresponding weights for the wider region are also correct for the City of Hayward, which may not be the case. No additional expert judgment was applied in the case of Hayward. We can better constrain structural system using more accurate data for occupancy class, number of stories, year built, and building material, but even in this case, we rely on the underlying structural systems and weights specified by Hazus.

There are also limitations in using Hazus vulnerability models to begin with. The Hazus models were not developed based on the Hayward building stock, and no additional analysis was done to see how well the Hazus vulnerability models describe Hayward structures. Beyond that, it is possible to have more specific typologies than Hazus (e.g., differentiating wood housing into 1 versus 2 stories). However, selecting to use Hazus vulnerability models is not a limitation in the inventory framework or methods proposed here, but rather is a study design choice that informs the required typology resolution of the inventory. For example, if a more detailed set of vulnerability models were used, there would be higher data demands (i.e., additional required building features) placed on the inventory.

In summary, of the five inventories, the *Best Estimate* inventory has the highest fidelity. This inventory incorporates all local data, which allows for several building features to be described using mostly collected data, including the occupancy class, year built, number of stories, number of units, and building material. This inventory improves fidelity relative to the *Synthesized Local* inventory by also incorporating national data sources to fill gaps and introducing specific data sources for occupancy classes that can be mischaracterized in the parcel data due to tax-exempt status, such as schools, universities, and government buildings. Furthermore, this inventory has the most complete and accurate baseline geometry (building footprints) because footprints corresponding to parking structures are removed (see Figure 10) and additional footprints are incorporated for cases where they are likely missing (see Figure 9g). While we do not have a perfect "ground truth", the *Best Estimate* inventory likely has higher accuracy than the nationally-derived inventories due to the high amount of collected data, and it has more complete coverage than the *Synthesized Local* inventory. Improved accuracy for many building features and the use of building material to constrain selection both improve the structural system assignment, even if building-level structural system data remains unavailable. Other fidelity limitations for the *Best Estimate* inventory include a reliance on estimated NSI data for replacement cost and building population.

In the absence of local data, the *Synthesized National* inventory represents an increase in fidelity relative to the *NSI Point* and *NSI Spatial Join* inventories because it incorporates additional HIFLD data to describe certain occupancy classes, avoids the bias in NSI building material, improves the classification of mobile homes, better approximates the number of units by incorporating census data, and adjusts the number of stories to remove unrealistically tall buildings. These changes all contribute to increased fidelity in the final inventory, which is evidenced by the *Synthesized National* inventory more closely approximating the *Best Estimate* inventory. Table 6 summarizes the percent agreement across different building inventories, demonstrating the benefits of the *Synthesized National* inventory relative to the *NSI Point* inventory. In terms of occupancy class, single-family (RES1) buildings show high agreement across all sources; however, comparisons among buildings other than single-family homes make the benefits of the *Synthesized National* inventory even more apparent. That being said, there are still fidelity limitations in the *Synthesized National* inventory, including low structural system accuracy, as described above, reliance on estimated NSI data for replacement cost and building population, and the use of NSI for year built, which remains low in accuracy because it is only available as a median at the block group level.

Table 6: Footprint-Level Building Material 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 Material: 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%

5.2. Implications for Seismic Risk

Discrepancies in building features are important for several reasons. First, when a study intends to focus on a specific subset of buildings (e.g., older concrete buildings or schools), correct identification of those buildings in the inventory is essential. Otherwise, errors or incorrect assumptions about the building features could lead to biases and invalidate the study. Second, since building features are commonly used to assign vulnerability models, the incorrect identification of building features can lead to significant errors in the resulting risk estimates. Finally, when tied to local demographics, accurate assessment of the building inventory is important to identify socio-economic factors that can magnify the impacts on the local community and its ability to recover.

In the case of Hayward, the impact of inventory differences on seismic risk estimates is demonstrated by evaluating the damage and loss from a large-magnitude (M7) earthquake on the Hayward Fault across the *NSI Point*, *Synthesized National*, and *Best Estimate* inventories. Figure 17 compares building-level mean loss ratios for each inventory under the same scenario earthquake [78]. The results reveal notable differences in overall loss estimates along with differences in the spatial distribution of losses across the city. Specifically, the seismic risk captured by the *Synthesized National* inventory (middle panel) closely approximates that of the *Best Estimate* inventory (right panel), while results for the *NSI Point* inventory (left panel) differ significantly, primarily due to its overestimation of masonry construction in residential buildings and its under-counting of mobile homes. In ongoing work, the authors are conducting a more systematic examination of how inventory development decisions propagate into seismic risk metrics and their associated uncertainties [79].

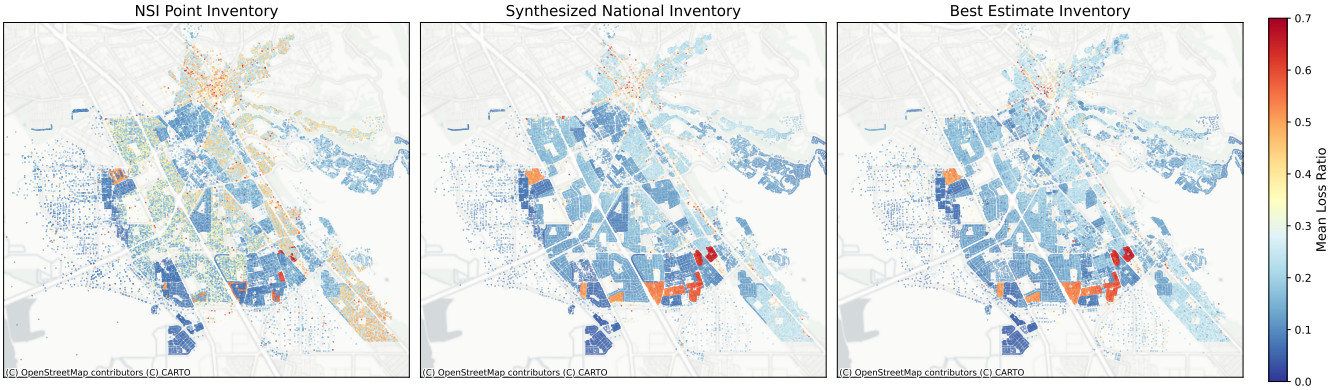

Figure 17: Comparison of the mean loss ratio in a scenario M7 earthquake on the Hayward Fault for the *NSI Point*, *Synthesized National*, and *Best Estimate* inventories.

6. Generalization to Other Contexts

While results and implications are shown in the context of seismic risk in Hayward, CA, the proposed terminology, framework, and methods are intended to generalize across hazards and locations. The terminology introduced in Section 2 applies broadly to describing and classifying inventory data, and the general framework steps shown in Figure 2 are relevant across applications regardless of input data sources, location, or hazard. Several of the specific methods developed to support these steps (highlighted in Figures 4 and 8) are also designed to be generalizable to other contexts.

6.1. Generalization to Other Hazards

The inventory synthesis framework and terminology are applicable for multiple types of hazards. While some building features like occupancy class, replacement cost, and population are broadly relevant, the importance of other building features varies by hazard. For example, roof shape is important for understanding hurricane risk, the presence of a basement matters in the context of flooding, while the structural system is important for earthquakes. Thus, while the framework and methods are generalizable, the required building features depend heavily on the hazard of interest and selected vulnerability models.

The importance of precise building location also varies by hazard. For hazards such as riverine or coastal flooding and storm surge, hazard intensity can change significantly over short distances, making accurate building location essential. In contrast, seismic hazard varies more gradually, meaning results are likely less sensitive to exact building location. As a result, the importance of building-level inventories (as opposed to aggregated inventories)

generally increases for hazards other than seismic risk. Aside from the *NSI Point* inventory, all Hayward inventories discussed in Section 5 use Hayward building footprints as their baseline geometry. Thus, building locations do not change across inventories, except when building footprints with no data are dropped. The *NSI Point* inventory has many differences in building locations, including unrealistic points located far from any building footprints, along transit lines, and more. Thus, if findings in Hayward are indicative of broader trends, synthesizing data into building footprints (rather than using raw NSI location) improves characterization of building location.

6.2. Generalization to Other Locations

The framework is also broadly applicable across different locations. Most explicitly, the national synthesis workflow (Figure 4) and the code to support it are designed for direct application across the United States, and preliminary work has adopted it for locations beyond Hayward [80]. Although some methods were informed by the Hayward context, the overall national synthesis workflow is expected to be broadly applicable. Some of the California-specific steps, like the modifications to structural system assignment, may not be relevant elsewhere and can easily be omitted. Since we have not tested the national synthesis workflow in many locations, it is possible that there are data inaccuracies or issues that do not arise in Hayward that may be relevant elsewhere. For example, since the baseline geometry is building footprints, if a certain community has incomplete building footprint data, this would be a limitation in the national synthesis workflow.

Beyond the direct implementation of the national synthesis workflow, several components of the framework transfer to other locations in the United States and beyond. The general framework is designed to be flexible across different inventory development applications and bring more structured and standardized thinking to inventory development. In addition, the methods described in Tables 3 and 4 (i.e., point-to-footprint and parcel-to-footprint) can also be adopted for other data sources because they are only dependent on the source geometry of the inputs, not the actual building features being described. The framework is also adaptable to different vulnerability modeling approaches. For instance, replacing the Hazus vulnerability models with the GEM taxonomy [81], which is broader, more flexible, and globally applicable, would change the building features required for vulnerability model assignment, but the overall synthesis framework would remain the same.

The proposed framework could also support the development of harmonized building inventories, which seek to describe the building inventory consistently across different regions despite differences in underlying data availability. Babič et al. (2025) identify two key challenges for harmonized exposure models: determining the optimal scale of analysis given available data detail in each region, and selecting structural and typological parameters appropriate for classifying buildings into fragility classes [49]. The inventory synthesis framework proposed in this study could help address both. The inventory design decisions discussed in Section 3 focus on selecting an appropriate spatial and typology resolution, which would need to be consistent across regions. Then, if a consistent baseline geometry can be defined across the whole study area, the inventory synthesis framework lays out steps for pulling input data from heterogeneous data sources (i.e., region-specific data). These data sources can be preprocessed and attributed to the baseline geometry using a method appropriate to each data source and the selected baseline geometry (i.e., point-to-footprint attribution, disaggregation from a census-level dataset, etc.). Standardized methods across region boundaries can be used to resolve disagreement, impute remaining gaps in the inventory, infer missing features, and map to features required for simulation. Furthermore, concurrent work focuses on uncertainty quantification in the building inventory, which would be particularly relevant when uncertainty is much higher in some regions than in others due to underlying data availability and accuracy.

Finally, this discussion focuses on data-rich contexts, where multiple data sources 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 very data-scarce environments. Studies from GEM and others recommend methods they found useful to create building inventories in such contexts [20, 23, 82].

7. Summary and Conclusions

This study outlines relevant terminology and proposes a systematic framework for synthesizing multiple data sources to develop a building inventory for the purpose of regional risk assessment. The framework can accommodate various input data types, which can be characterized by their 1) data provenance, which describes whether the data is collected, estimated, or of uncertain provenance, 2) source geometry, which describes both the geometry and scope of what the data represents, and 3) building features, which refer to the structural, use-related, value, and socio-economic characteristics described in the inventory data. Depending on input data availability, the framework can be used to develop inventories with various levels of fidelity. Fidelity is characterized by the

spatial resolution, typology resolution, building feature accuracy, and completeness of the final inventory. A key conceptual contribution of this framework is the distinction between inventory resolution and inventory fidelity: increasing spatial or typology resolution only improves fidelity when additional information is incorporated, not simply when data is disaggregated or sampled. The proposed inventory synthesis framework involves preprocessing data, attributing sources to a baseline geometry, selecting values for features with disagreement, filling gaps in building feature values through imputation or inference, and mapping to features required for simulation.

Specific examples of how the framework can be used are demonstrated for the case study of building inventories for a regional seismic risk assessment in Hayward, CA. Three specific workflows based on the general inventory synthesis framework are shown for Hayward: the national synthesis workflow, which uses only nationally-available data, the local synthesis workflow, which uses only local data from Hayward, and the best estimate workflow, which incorporates all available input data to create an inventory for the city. As part of these workflows, several specific methods were developed for 1) general point-to-footprint attribution, 2) parcel-to-footprint attribution, 3) synthesis of NSI and HIFLD data, 4) a method to leverage census data to improve the estimation of the number of housing units in a given building. In addition, a complete implementation of the national synthesis workflow is available for use in other locations in the United States with very little modification.

The Hayward case study illustrates how the choice of input data sources and inventory synthesis methods change the resulting inventory composition. A comparison of results for five building inventories, all nominally representing Hayward circa 2023, demonstrates the significant differences that can arise in building features such as occupancy class, number of units, and building material. When propagated through a scenario M7 earthquake, differences in the building inventory composition result in large differences in the quantity and distribution of seismic risk. Ongoing work by the authors aims to quantify seismic risk impacts and uncertainty more comprehensively.

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

Different data sources and data synthesis methods lead to significant differences in the resulting building inventories. This study reveals that the choice of input data sources and inventory synthesis methods can result in significant differences in the quantity of buildings and distribution of building features across the city of Hayward. For example, the selection of which footprint data source will be used as the baseline geometry alone can affect the inventory significantly, with footprint sources like OpenStreetMap causing up to 24% of the buildings being missing, with corresponding losses in population and replacement cost. This demonstrates that input data source selection, particularly that of the baseline geometry, should be treated as a substantive decision rather than a logistical one. Even when a robust baseline geometry is used, selected input data sources and synthesis methods still result in significant differences. Aggregate metrics, such as total building count or population, vary moderately across inventories; however, more substantial differences appear when looking at more detailed metrics, such as the distribution of occupancy classes or building material. When propagated through to building-level loss for a scenario earthquake, building inventory differences translate to large differences in the distribution of loss across the city. To help contextualize and make regional-level assessments more usable for risk mitigation, more standardized and systematic approaches towards inventory development 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 feature (e.g., occupancy classes), which may cause bias 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 and building features into a single, consistent building inventory. Synthesizing multiple data sources enables the final inventory to leverage the strengths of different inputs, such as the complete coverage of national inventories like NSI, the accuracy of occupancy-class specific data from HIFLD, and the local specificity of tax and address data. Depending on the available input data sources, their completeness, and the provenance and accuracy of the building features they describe, this process can lead to a higher-fidelity final inventory than any individual data source can provide. An additional benefit of this process is that comparing multiple sources can surface systematic inaccuracies that would be invisible when relying on any single source alone; between-source disagreements flag potential biases in the underlying data and can prompt manual checks and careful source

prioritization that further improve inventory quality. For data synthesis to deliver these benefits, however, methods must be selected carefully to avoid inadvertently dropping data or introducing additional bias. Traditional spatial merging techniques often fall short in this regard because inventory data frequently has approximate geometries and lacks one-to-one mapping across sources. The point-to-footprint and parcel-to-footprint attribution methods proposed in this study are designed to address these challenges and move towards more standardized, reproducible approaches to data synthesis.

Synthesizing nationally available and locally obtained data leads to the best estimate inventory; however, if only national data sources are used, there are simple steps that can help reduce bias, address common issues, and improve fidelity. This study develops a best estimate inventory by synthesizing both national and local data sources for Hayward. Whereas standardized methods are proposed for common national data sources, integration of local tax and address data requires considerable effort because specific processes are needed to obtain, preprocess, and synthesize data that follows the specific schema used by each local jurisdiction. The national synthesis workflow, summarized in Figure 4 and publicly available on GitHub, presents a straightforward way to improve on the commonly-used NSI data without requiring local data efforts, and has already been applied in other locations beyond Hayward. The workflow addresses several common errors and biases observed in Hayward, which are expected to 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, 5) significantly over-counting the presence of masonry in residential buildings, and 6) inaccurately representing educational and emergency response structures. There are several remaining limitations in the fidelity of the national workflow as it currently stands: structural system is sampled from the Hazus mapping scheme tables (see Figure 7), and likely has low building-level accuracy. In addition, only a median year-built estimate is available, and the final inventory relies on NSI estimates of replacement cost and population. Reducing the identified errors and biases through careful national data synthesis results in a higher-fidelity building inventory, but these limitations represent a current ceiling on the fidelity achievable without collected data.

There are several inventory development concepts that remain open questions 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. Beyond just urban growth, inventory temporality is also important on smaller time scales, which is not typically accounted for. For example, in university towns or areas with heavy tourism, seasonal population swings can meaningfully shift the amount and distribution of risk.

Future work should also quantify inventory uncertainty. This study resolves within- and between-source disagreements and fills gaps using imputation and inference by selecting single instances of building feature values. While this approach is practical, synthesizing multiple data sources and handling gaps in this way can obscure uncertainty in the inventory. The framework proposed in this study could be used to support a more robust approach towards uncertainty in the inventory data, where multiple inventory realizations are propagated through a risk assessment. Ongoing work by the authors addresses this topic [79].

Another key area for future work is improving structural system classification. In many locations around the country and the world, structural system information at the building level is not available. Thus, improved methods for capturing the structural system at the building level with confidence, through the use of local expert judgment or validation, targeted visual surveys for key buildings, or leveraging computer vision could be helpful in improving building-level structural system classification. Progress in this area would meaningfully improve the ceiling on inventory fidelity that currently constrains all inventories discussed in this study.

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 terminology, open-source workflows, and standardized methods for synthesizing inventory data helps reveal potential biases, understand the impact of inventory development decisions, and ultimately move towards higher-fidelity building inventories that more faithfully capture the built environment and the communities within it.

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 repository is intended not only to support reproducibility, but also to serve as a general-purpose tool for similar exposure modeling workflows in other contexts. More broadly, the repository is intended to promote shared, transparent workflows and more standardized approaches to inventory data synthesis.

The core Python functions are designed to support inventory generation beyond the Hayward case study with minimal modification, and we encourage their application and extension in new settings. In particular, the Synthesized National workflow can be run for any area in the United States with minimal modification. In addition, individual components of the repository, including scripts for attributing points to building footprints, attributing parcel polygons to building footprints, resolving within- and between-source disagreement, and plotting data geospatially, can be used as a template for other contexts.

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 opinions, 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 Claude and ChatGPT to edit 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 inventory data correctly represents the real-world building inventory, regardless of spatial or typology resolution. Different building features can have different levels of accuracy, even within a single data source, as some building features are better documented or easier to capture than others.

Baseline Geometry: The selected spatial representation used to describe the final inventory, which should be appropriate to the study's targeted spatial resolution. In the proposed framework, data synthesis is enabled by attributing all input data sources to a consistent baseline geometry source. The framework in this study assumes single-building polygons are used as the baseline geometry, with footprints being the primary example of single-building polygons. Selection of a baseline geometry source is a critical step, as gaps and errors in the baseline geometry source will directly result in gaps and errors in the final synthesized building inventory.

Between-Source Disagreement: When different input data sources report conflicting information about a given building feature of a single building footprint.

Bounding Geometry: Bounding geometry describes a dataset with multi-building polygon source geometry that is used to assist in the process of attributing data to footprints. Its main purpose is to 1) improve computational efficiency by constraining the number of distance calculations from points to footprints to those in the same (or adjacent) bounding geometry polygons, and 2) provide reasonable distance limits for point-to-footprint attribution. The bounding geometry polygons should be non-overlapping and have complete coverage across the study area.

Building Features: The descriptive characteristics of a building provided by a data source, other than its source geometry. These can include use-related information (e.g., occupancy class), building value (e.g., replacement cost), structural attributes (e.g., number of stories, structural system), and socio-economic characteristics (e.g., number of residents, income, tenure).

Building Material: The primary structural material of a building. This concept is also referred to as 'building type' in the National Structure Inventory (NSI) and Hazus 6.1 Inventory. The building materials adopted in this study follow the five general building materials in the Hazus 6.1 Inventory: wood, steel, concrete, masonry, and manufactured housing.

Collected Inventory Data: Building inventory data that has been directly observed or recorded through administrative processes, surveys, or measurements, such as property tax assessor records or building permit data. These data sources are often considered closer to ground truth, but they may be incomplete, inconsistently defined across jurisdictions, or missing attributes relevant for hazard analysis.

Completely Missing Building Feature: A building feature is completely missing when it is not available for any building in the inventory.

Estimated Inventory Data: Building inventory data that has been modeled, inferred, or estimated from other information using rules, statistical models, or remote sensing techniques. Such approaches are commonly used to create spatially complete data at large geographic scales, but they typically have lower inventory accuracy than is typical of collected inventory data.

Fidelity: The extent to which the inventory data represents the real building stock in terms of accuracy, completeness, spatial resolution, and typology resolution. Creating higher-fidelity inventories involves increasing the spatial and/or typology resolution, while 1) maintaining at least as much accuracy in each building feature as the lower-resolution inventory and 2) ensuring that the data is complete, meaning that buildings are not missing from the inventory.

Imputation: The process of filling in partially missing building features using statistical methods based on patterns found in existing data. Imputation of a given building feature can be done using only available data for that building feature, though it can also use additional assumptions and judgment.

Inference: The process of filling in completely missing building features through the use of external models or reasonable assumptions based on available building features.

Uncertain-Provenance Data: Building inventory data for which there is insufficient information to determine whether the individual buildings and/or building features being described in the data source are collected or estimated.

Partially Missing Building Feature: A building feature is partially missing when it is available for a subset of buildings in the inventory. Partially missing information can either be random (gaps occur without a predictable pattern in otherwise available features), or systematic (gaps occur in a predictable way due to factors such as data collection practices).

Occupancy Class: The primary use and/or 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. 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 B.

Source Geometry: The spatial representation of a given data source. The source geometry describes both 1) the geometry (i.e., whether it is comprised of points or polygons) and 2) the scope of what the data represents (i.e., whether a single point/polygon in the data source represents a single building, a single unit, multiple buildings, or a mix of these). Examples include single-building polygons (e.g., building footprints), multi-building polygons (e.g., census tracts or college campuses), single-unit points (e.g., address point data), and more. Please see the list provided in Section 2 for a comprehensive list of source geometries.

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

Structural System: A classification describing the primary structural system of a building, also referred to as structure type. Structural system is a more specific classification than building material (i.e., if the building material is 'steel', the structural system could be 'steel moment frame', 'steel braced frame,' etc.). The term structural system, 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.

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. The targeted typology resolution determines which building features must be captured in the inventory to support vulnerability model selection.

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. Extended Occupancy Class Information

Most of the possible values for occupancy class in this study 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 B.7: 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

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 B.8: 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	Local data specifies mixed-use buildings	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 C. 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 was 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 Hazus 6.1 [9], it is not broadly available to researchers. NSI is complete in that it does not have any gaps or missing features; however, some building features appear to be more accurate 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 a vulnerability model, 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 complete, NSI has notable limitations, especially for users relying only on the publicly available fields.
- **USA Structures**: USA Structures is a single-building polygon source describing all buildings 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. At the time of submission, 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.). Its scope does not aim to be comprehensive for all buildings, like NSI and USA Structures, but instead geographically represents 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 occupancy class, 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 26, 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 [83].

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 because they represent fundamentally different things. 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, an address point could represent 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 may require a significant amount of preprocessing to extract helpful building inventory information.

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. 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 at the time of submission, it was most recently updated in 2025.

- **OpenStreetMap**: OpenStreetMap is a global building footprint database (single-building polygons) which is open-source and community-developed, thus differentiating it from the other global footprint sources. As a result, 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 an industry consortium launched in 2022 under the Linux Foundation 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 data sources, including OpenStreetMap, Esri Community Maps, Microsoft Building Footprints, Google Open Buildings, and others [67]. The synthesis process first prioritizes community-contributed data, then supplements the rest with the best machine-learning-based data available. This dataset is actively being developed and at the time of submission, was most recently updated in 2025.

Appendix D. Collected and Modeled Features across Inventories

In Section 2, three terms are introduced to characterize data provenance: *collected inventory data*, *estimated inventory data*, and *uncertain-provenance inventory data*. These terms can be used to describe the origin of data sources, which can be helpful in understanding their likely level of accuracy.

The five inventories presented in Section 5 are each developed using different input data sources and synthesis steps, as described in Section 4. To summarize this information in a simplified and concise manner, Figure D.18 shows the data provenance of building features across the five inventories.

Two important notes apply to these classifications. First, the three categories are broad, and a range of accuracy and completeness can fall within each category. Second, almost all building features in the three synthesized inventories draw from multiple sources, with some information being imputed or taken from secondary sources, as described in Section 4. Because the national, local, and best estimate inventory synthesis workflows track the origin of each building and building feature, the provenance of each can be identified, distinguishing the resulting data from uncertain-provenance data. The classifications shown in Figure D.18 represent the data provenance for the majority of buildings in each inventory for a given building feature.

Figure D.18 is meant to provide an overview of data provenance, and it is not meant to be a comprehensive description of the inventories. Please refer to Figures 4 and 8 for more detailed information.

	NSI Point	NSI Spatial Join	Synthesized National	Synthesized Local	Best Estimate
Building Location	NSI Point Locations	Hayward Footprint Polygons (Dropped if no data)	Hayward Footprint Polygons (Dropped if no data)	Hayward Footprint Polygons (Dropped if no data)	Improved Hayward Footprint Polygons (Dropped if no data)
Occupancy Class	Uncertain-Provenance (NSI)	Uncertain-Provenance (NSI)	Uncertain-Provenance (NSI) + Collected for certain occupancy classes (HIFLD)	Collected (Address/Parcel)	Collected (Address/Parcel) + Collected for certain occupancy classes (HIFLD)
Year Built	Estimated (NSI Median Year)	Estimated (NSI Median Year)	Estimated (NSI Median Year)	Collected (Parcel)	Collected (Parcel)
Number of Stories	Uncertain-Provenance (NSI)	Uncertain-Provenance (NSI)	Uncertain-Provenance (NSI), with correction for tall buidlings	Collected (Parcel)	Collected (Parcel)
Number of Units	NA	NA	Estimated (using Census Units and NSI Population)	Collected (Address Points per Footprint)	Collected (Address Points per Footprint)
Building Material	Uncertain-Provenance (NSI)	Uncertain-Provenance (NSI)	Mapped from Structural System	Collected (Parcel)	Collected (Parcel)
Structural System	Estimated (Inferred using Hazus Tables, contrained by NSI Building Material)	Estimated (Inferred using Hazus Tables, contrained by NSI Building Material)	Estimated (Inferred using Hazus Tables)	Estimated (Inferred using Hazus Tables, contrained by Parcel Building Material)	Estimated (Inferred using Hazus Tables, contrained by Parcel Building Material)
Replacement Cost	Estimated (NSI)	Estimated (NSI)	Estimated (NSI) + Estimated (Hazus) when needed	Estimated (NSI) + Estimated (Hazus) when needed	Estimated (NSI) + Estimated (Hazus) when needed
Population	Estimated (NSI)	Estimated (NSI)	Estimated (NSI)	Estimated (NSI)	Estimated (NSI)

LEGEND
**Please note, the three synthesized inventories have some data that is imputed or taken from secondary data sources. The classification shown in this table corresponds to the data provenance of the majority of buildings in the inventory for the given building feature.
Collected: Data that has been directly observed or recorded through administrative processes, surveys, or measurements, such as property tax assessor records or building permit data
Estimated: Data that has been modeled, inferred, or estimated from other information using rules, statistical models, or remote sensing techniques.
Uncertain-Provenance: Data for which there is insufficient information to determine whether individual buildings and/or building features are collected or estimated

Figure D.18: Data Provenance for Majority of Buildings, by Building Feature and Inventory.

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] P. Gamba, Global exposure database: Scientific features, Global Earthquake Model (GEM) Foundation, Pavia, Italy 710 (2014).

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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

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

[38] 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.

[39] 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.

[40] 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).

[41] 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).

[42] 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.

[43] 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).

[44] 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).

[45] 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.

[46] 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.

[47] 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.

[48] B. R. Ellingwood, C. Harvey, P. Gardoni, P. Walter Gillis, J. W. van de Lindt, N. Wang, The Centerville Virtual Community: a fully integrated decision model of interacting physical and social infrastructure systems 1 (3-4) (2016) 95–107. doi:10.1080/23789689.2016.1255000.

[49] A. Babič, M. Polese, G. Tocchi, M. Faravelli, B. Borzi, M. Dolšek, A framework for harmonized cross-border seismic risk assessment, Bulletin of Earthquake Engineering 23 (6) (2025) 2421–2449.

[50] M. Polese, G. Tocchi, A. Babič, M. Dolšek, M. Faravelli, D. Quaroni, B. Borzi, N. Rebora, D. Ottonelli, S. Wernhart, et al., Multi-risk assessment in transboundary areas: A framework for harmonized evaluation considering seismic and flood risks, International Journal of Disaster Risk Reduction 101 (2024) 104275.

[51] 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.

[52] 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.

[53] 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.

[54] 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).

[55] C. K. Huyck, Z. Hu, M. Eguchi, G. Esquivias, P. Amyx, K. Smith, C. Jordan, Characterizing uncertainty of general building stock exposure data, Earthquake Spectra 38 (3) (2022) 2008–2025.

[56] 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.

[57] P. Kalakonas, V. Silva, A. Mouyiannou, A. Rao, Exploring the impact of epistemic uncertainty on a regional probabilistic seismic risk assessment model, Natural Hazards 104 (1) (2020) 997–1020.

[58] A. Pollack, J. Benedict, M. Deb, J. Doss-Gollin, D. Judi, W. Lehman, N. Lutz, C. Reesman, E. Sarazen, Y. Son, V. Srikrishnan, N. Sun, K. Keller, Unrefined national building inventories can mislead risk assessments and decisions, SSRN Electronic Journal (October 2025). doi:10.2139/ssrn.5575271.

[59] 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.

[60] 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.

[61] A. Babič, J. Žižmond, M. Dolšek, Bias in the estimation of seismic risk for municipal building stocks due to limited data, Buildings 13 (9) (2023) 2245.

[62] I. E. Bal, J. J. Bommer, P. J. Stafford, H. Crowley, R. Pinho, The influence of geographical resolution of urban exposure data in an earthquake loss model for istanbul, Earthquake Spectra 26 (3) (2010) 619–634.

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

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

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

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

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

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

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

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

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

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

[73] A. Mignan, Introduction to Catastrophe Risk Modelling: A Physics-based Approach, Cambridge University Press, Cambridge, UK, 2025. doi:10.1017/9781009437370.

[74] Z. Song, W. Jiang, S.-r. Yi, J. Zhang, Reliable building inventory imputation for regional-scale risk assessment: An uncertainty-guided framework using spatially-enhanced transformers, Reliability Engineering & System Safety 274 (2026) 112436.

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

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

[77] 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).

[78] F. McKenna, S. Gavrilovic, J. Zhao, K. Zhong, A. Zsarnoczay, B. Cetiner, S. Naeimi, S. ri Yi, A. B. Satish, P. Arduino, W. Elhaddad, NHERI-SimCenter/R2DTool: Version 5.3.0, zenodo (2025). doi:10.5281/zenodo.14800304.
URL https://doi.org/10.5281/zenodo.14800304

[79] M. Lochhead, A. Huynh, A. Zsarnoczay, G. G. Deierlein, Uncertainty and bias in footprint-level building inventories and their impact on regional seismic risk assessment, in: Proceedings of the 13th National Conference on Earthquake Engineering, 2026, forthcoming.

[80] S. Meiler, N. Blagojevic, M. Lochhead, J. W. Baker, Hurricane recovery potential, presented at the Symposium on Tropical Cyclone Risk in a Changing Climate, Tampa, Florida, June 2025 (2025).

[81] S. Brzev, C. Scawthorn, A. W. Charleson, L. Allen, M. Greene, K. Jaiswal, V. Silva, GEM Building Taxonomy (Version 2.0), Tech. Rep. 2013-02, GEM Foundation, publication Title: GEM Technical Report (2013).
URL https://pubs.usgs.gov/publication/70058718

[82] 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.

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