4.4. Data Fusion

From the public sources mentioned above, spatial and geo-point data categories are available and can be integrated to enhance the completeness and comprehensiveness of the building inventory. Previously, challenges in data fusion—such as matching instances across different sources (*e.g.*, determining which FEMA polygon corresponds to a Microsoft entity)—were discussed, primarily caused by spatial dislocation errors or missing records. To address these issues, imputed structural bounding boxes were introduced, enabling a two-phase data fusion strategy to be devised. The first phase involves fusing spatial geo-data (*i.e.*, FEMA and Microsoft polygons), as summarized in Algorithm 2. The core logic of the algorithm is straightforward: if a FEMA and a Microsoft polygon have an Intersection over Union (IoU) greater than 0.5—forming a symmetric and exclusive match—they are considered a valid pair. Examples of such matches are illustrated in Fig.12a. The statistics for this 'perfect' matching scenario are presented as 'Category I' in Table 5, with values ranging from approximately 0.5 to 0.7 across the studied counties. If one or more FEMA and Microsoft polygons intersect but with a lower IoU, the matching may still be one-to-one (see Fig.12b) or may involve a chain of multiple intersecting polygons (see Fig.12c). The latter case is often attributed to FEMA's limited precision in geo-locating or estimating building footprints. In these cases, the imputed bounding boxes support polygon matching resolution, as further described in Algorithm 2. Specifically, if a FEMA and a Microsoft polygon both exhibit the highest IoU with the same bounding box, they are treated as a match. The statistics for these cases are reported under 'Category II' in Table 5.

Up to this point, and given the high precision of Microsoft polygons, the imputed bounding boxes generally align well with the Microsoft polygon boundaries. However, Microsoft's delineation errors can occasionally be substantial, in which case the imputed bounding boxes help correct these discrepancies. Examples of such instances are shown in Fig.12d, where Microsoft had merged various structures in one in error (Category III). Algorithm 2 accounts for these scenarios, enabling the fusion process to rectify Microsoft polygons when both the FEMA data and the imputed bounding boxes concur on an error. The remaining Microsoft polygons that do not match any FEMA polygons are still retained in the inventory but lack FEMA-derived features, ranging from 10% to 40% of the available Microsoft data per Table 5, *i.e.*, 1− Categories I and II. Despite the advantages of the imputed boundaries discussed above, Categories V and VII (as defined in Algorithm 2) play a particularly significant role. First, unmatched FEMA polygons can be validated against the imputed bounding boxes to confirm whether they correspond to actual structures. FEMA polygons lacking such alignment can be discarded as potential false positives (see Fig.12e). Second, imputed bounding boxes that are not matched by either Microsoft or FEMA polygons contribute additional, previously unrecognized structures to the inventory, thus enhancing its completeness (see Fig. 12f. However, similar to the unmatched Microsoft polygons, these new imputed structures do not have associated features or known building footprint boundaries—issues which are addressed later in this section. The statistics on these additional structures—those not captured by Microsoft polygons—are summarized in Table 5, with potential gains in structural coverage reaching up to 15% relative to using Microsoft polygons alone.

With the geo-spatial data fused, the next step is geo-point data matching—*i.e.*, fusing NSI features with Microsoft and FEMA data. This procedure is outlined in Algorithm 3. The approach is straightforward: each fused polygon from Algorithm 2 may consist of a Microsoft polygon alone, a FEMA polygon alone, both combined, or a single imputed bounding box. In contrast to a vanilla matching method, where an NSI point is considered matched if it falls within a Microsoft polygon, the enhanced approach also allows matching with FEMA polygons—either

Algorithm 2 Microsoft–FEMA Geo-Spatial Data Fusion

Require: Microsoft polygons M, FEMA entries F, Imputed bounding boxes O

Ensure: Fused Microsoft–FEMA data: $M - F$

Find intersecting FEMA and Microsoft polygons
Keep non-intersecting FEMA and Microsoft polygons for later evaluation
for all intersecting M–F pairs **do**
  if IoU$(M, F) > 0.5$ **then**
    Merge FEMA features with Microsoft polygon
    Add to final fusion $M - F$ *// Category I*
  else if FEMA and Microsoft share the same matching imputed box from O **then**
    Match Microsoft and FEMA with highest IoU
    Merge FEMA features with Microsoft polygon
    Add to $M - F$ *// Category II*
  else if No shared imputed box, but all FEMA-matched boxes lie within the Microsoft polygon **then**
    if Only one FEMA-Microsoft pair exists **then**
      Keep Microsoft polygon in fusion
    else
      Discard Microsoft polygon
      Retain all intersecting FEMA polygons
    end if
    Add to $M - F$ *// Category III*
  else
    Discard intersecting FEMA polygons
    Retain Microsoft polygons
    Add to $M - F$ *// Category IV*
  end if
end for
for all non-intersecting FEMA polygons **do**
  if FEMA matches an imputed box (centroids within each other) **then**
    Add FEMA to $M - F$ *// Category V*
  else
    Discard FEMA
  end if
end for
for all non-intersecting Microsoft polygons **do**
  Add Microsoft polygon to $M - F$ *// Category VI*
end for
for all Imputed boxes $o \in O$ **do**
  if No M or F centroids lie within o, and o's centroid not within any M or F **then**
    Add o to $M - F$ as a geometry with no features *// Category VII*
  end if
end for
return Final fused dataset $M - F$

Fig. 10. Data Fusion Step I: fusing geo-spatial data.

Algorithm 3 NSI Matching with FEMA-Microsoft-Bounding Boxes Fusion

Input: A matched polygon P derived from one of:
  • Microsoft polygon M and FEMA polygon F
  • FEMA polygon F only
  • Bounding box B only
Input: NSI dataset with geo-points n
Output: Upgraded matched polygons with appended NSI attributes
for each geo-point n in NSI **do**
  if P includes both M and F **then**
    if $n \in M$ **or** $n \in F$ **then**
      Mark n as matched
      Append attributes of n to P
    end if
  else if P includes only F **then**
    if $n \in F$ **then**
      Mark n as matched
      Append attributes of n to P
    end if
  else if P includes only B **then**
    if $n \in B$ **then**
      Mark n as matched
      Append attributes of n to P
    end if
  end if
end for
return Upgraded matched polygons

Fig. 11. Data Fusion Step II: Fusing matched polygons with NSI.

independently or alongside Microsoft polygons (see Fig.13a)—as well as with imputed bounding boxes that encapsulate NSI points but are not associated with either FEMA or Microsoft polygons (see Fig.13b). This extended matching capability, enabled by Algorithm 3, demonstrates how the fusion process improves the alignment between spatial and point-based data. It is important to note that NSI points are matched to imputed bounding boxes only if those boxes are not linked to any FEMA or Microsoft polygons. Otherwise, allowing matches to bounding boxes overlapping existing FEMA or Microsoft polygons may result in erroneous associations due to spatial overlapping. Additionally, if multiple NSI points are matched to a single footprint boundary—which, as discussed, may occur in the case of multi-family structures or commercial

settings (see Fig. 5)—all associated point attributes are retained and linked to that boundary.

Table 4
Fusion Step-I statistics, noted categories are defined per Algorithm 2.

State	Category I (% *)	Category II (%)	Unmatched (%)		Category V**	Category VII †	NSI Matching	
			Microsoft	FEMA			Initial	Fusion
Delaware	64.3	30.6	5.1	9.4	6.4	3.4	87.8	94.1
Maine	49.0	14.1	36.9	18.8	9.8	5.3	86.7	88.7
Minnesota	76.7	8.9	14.3	6.2	2.8	2.4	89.9	93.7
Nebraska	88.4	3.3	8.3	6.6	4.4	3.2	93.8	97.5
North Carolina	56.4	12.1	31.3	8.5	5.0	3.9	91.3	96.2
Oregon	67.4	25.6	7.0	4.5	2.6	3.4	92.6	96.1
Texas	53.4	8.1	3.8	7.5	12.3	3.7	81.8	95.2

* Ratio is calculated over the initial Microsoft polygon population.
** New footprints added by FEMA fusion.
† New footprints added by Imputed bounding boxes fusion.
‡ Overall footprint increase ratio compared to Microsoft footprints alone.

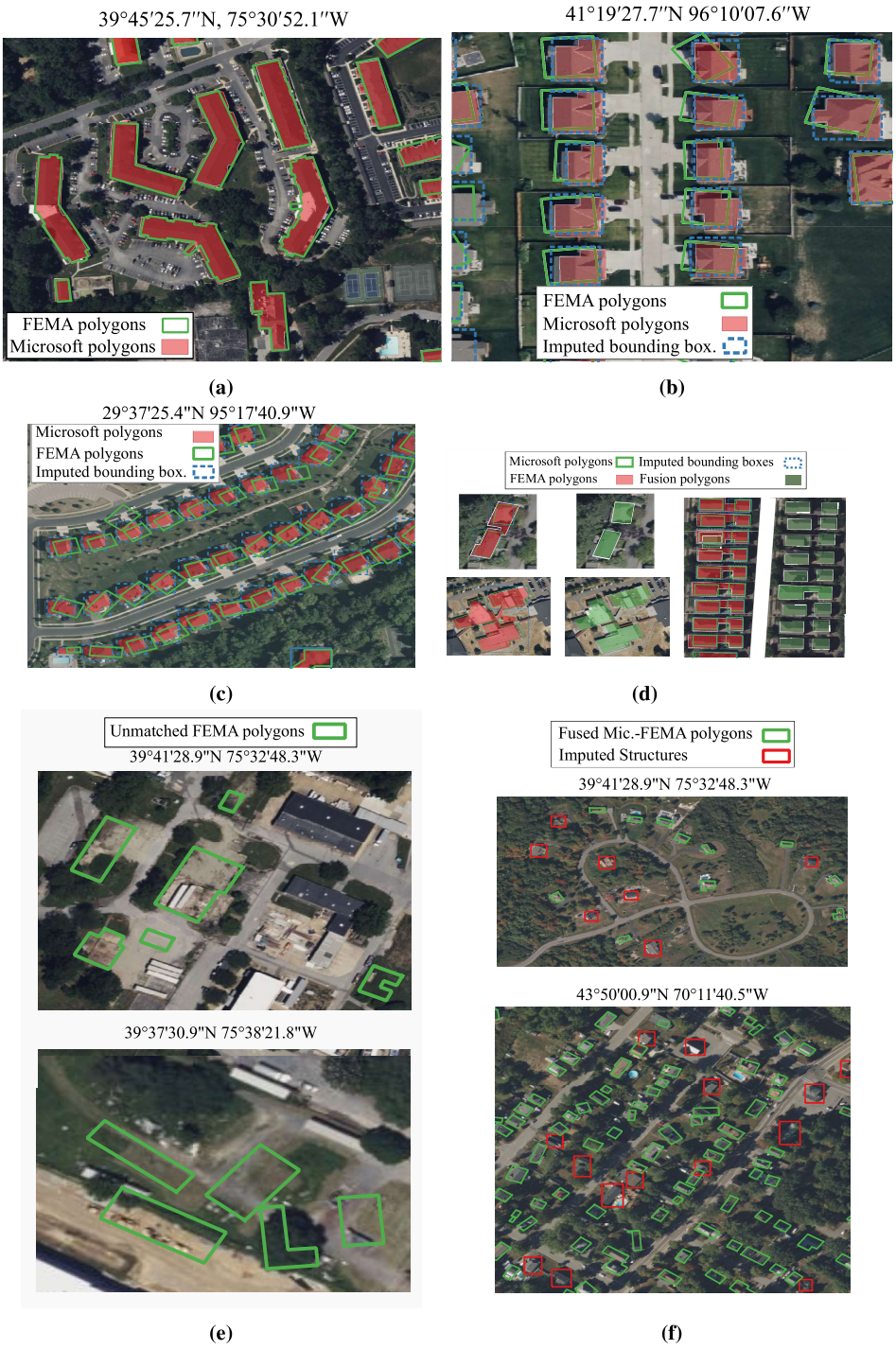

Fig. 12. Exemplar Algorithm 2's outputs: a) Category-I's matching, b) Category-II's (one-to-one) matching, c) Category-II's (multiple), d) Microsoft correction with FEMA and imputed bounding boxes, e) Discarded Category-IV's FEMA footprints with the help of (lack) of imputed boxes, f) Category-V's added structures through the imputed bounding boxes.

Fig. 13. Algorithm 3's examples; a) additional NSI and FEMA-Microsoft match through the fusion process, b) reviving unmatched NSI points with the imputed Bounding boxes.