
Shenglong Chen*, Yoshiki Ogawa**, Yoshihide Sekimoto***

Abstract: Traditional evaluation metrics for footprint extraction depend on manually generated ground truth. Besides, certain areas lack authoritative reference data, and intrinsic methods are confined in limited regional applicability. Therefore, the study introduces a two-level approach. Initially, population and land use data serve as a proxy for grid-level completeness assessment. Secondly, an enhanced two-way area overlap method is employed to match extracted footprints with reference buildings, yielding a comprehensive evaluation of global and semantic accuracy. Validation conducted in Hyogo Prefecture and Numazu City demonstrate an enhancement in grid classification accuracy and completeness correlation by 2.6% and 0.53. Furthermore, the optimized matching method achieves an impressive 99% accuracy in semantic matching, displaying exceptional efficiency and robustness in multi-scale matching. This study presents an exhaustive and efficient solution for establishing a comprehensive large-scale building extraction evaluation system.

Keywords: Footprint quality evaluation, Completeness assessment, Proxy data, Building matching.

1. Introduction

Recent studies have investigated using deep learning technologies to automatically extract large-scale building footprints from remote sensing images (Chen et al., 2023). However, challenges related to reliability, accuracy, and usability still arise because of building feature diversity, sensor performance limitations, and so on. Consequently, evaluating extraction performance has become a crucial aspect of these methods.

Numerous studies have examined extraction results, mostly focusing on traditional accuracy assessments for image segmentation (Dey and Awrangjeb, 2020). However, this approach is limited by its dependence on reference data quality, often relying on manually created ground truth for assessment. The evaluation results can significantly differ based on the selected sampling area, potentially not representing overall extraction quality adequately. A viable solution involves maximizing the utilization of existing reference building data. Ensuring correspondence between extracted footprints and real-world objects requires geographic matching, presenting two significant challenges: discrepancies in position due to off-nadir angle in satellite images and complex semantic variations from chronological and detail disparities (Wang et al., 2022). Moreover, matching a wide range of footprints places high computational efficiency demands on the algorithm.

In certain developing regions, accessing reliable reference data poses challenges, particularly in evaluating aspects like positional accuracy and shape similarity. While some studies suggest utilizing indicators like building density and count as proxies for completeness, these quantitative relationships might not universally apply beyond specific target regions (Tian et al., 2019). Alternative proxy data, such as population statistics and land cover information, have been proposed for completeness assessment (Zhang et al., 2022). However, relying solely on population data disregards sparsely populated regions such as rural or industrial areas.

This study introduces a two-level approach, designed to evaluate large-scale footprint extraction results. Firstly, a grid-based approach utilizes WorldPop population and Esri land cover data as proxies to assess footprint completeness (Zhou et al., 2022b). Secondly, an object-based method employs an optimized two-way area overlay (TWAO) technique to identify matches between

* GISA member College of Civil Engineering, The University of Tokyo chen-sl@csis.u-tokyo.ac.jp
** GISA member Center for Spatial Information Science, The University of Tokyo ogawa@csis.u-tokyo.ac.jp
*** GISA member Center for Spatial Information Science, The University of Tokyo sekimoto@iis.u-tokyo.ac.jp
extraction and reference data. (Fan et al., 2014). This yields global accuracy metrics and intricate semantic relationships. Finally, the methodology’s efficacy is demonstrated through evaluations in Hyogo Prefecture, Japan using the grid-based approach, and in Numazu City using the object-based approach.

2. Related works

2.1 Footprint completeness assessment

To assess the completeness of footprint data, quantitative analysis is utilized to quantify various factors within a given region, by comparing it with a reference dataset (Zhou et al., 2022a). To tackle the challenge of insufficient building datasets, scholars have suggested using machine learning techniques or crowdsourcing methods to deduce or acquire reference footprint data. Herfort et al. (2022) applied a random forest regression to estimate building completeness in a megalopolis with 13,189 locations. However, training a machine learning-based model requires a substantial volume of training data and the fine-tuning of numerous hyperparameters and model architectures. Ullah et al. (2023) adopted a gamified crowdsourcing approach and engaged volunteers through the mobile application MapSwipe to classify building completeness. Nevertheless, there’s room for improvement in both the quantity and quality of volunteer contributions. Moreover, several studies have employed building-related data as substitutes for the reference data (Zhou et al., 2022b). The availability of open-source population and land cover products with enhanced global resolution and precision has opened up new avenues for exploring innovative evaluation approaches.

2.2 Building-unit quality evaluation

Evaluation of the extraction involves two main steps: matching the results with the reference data and computing the corresponding quality metrics. The objective of building matching is to establish correlations between identical footprint characteristics from diverse sources. For polygons, the commonly used TWAO method accurately matches individual target positions but lacks precision in simultaneous matching of multiple targets (Memduhoglu and Basaraner, 2022). To enhance accuracy of complex semantic relations, researchers use techniques like minimum bounding rectangle combinatorial optimization (MBRCO) and relaxed labeling (Liu et al., 2023). However, these methods require exploring combinations and multiple iterations significantly increasing matching time. When considering evaluation metrics, Zeng et al. (2013) introduced three key aspects: matching rate, shape similarity, and positional accuracy. The matching rate, a crucial and commonly used measure, quantifies detection rate (recall) and accuracy (precision) based on building statistics (Chen et al., 2023). Shape similarity and positional accuracy assess quality through geometric criteria, relying on accurate reference data (Huang et al., 2021). This study primarily focused on the matching rate due to constraints in footprint extraction methods and available data.

3. Methodology

3.1 Study area and data

In Figure 1c, Hyogo Prefecture and Numazu City were chosen as our study areas. The purpose of Hyogo Prefecture was to test the grid-level method’s completeness using proxy data. Conversely, Numazu City was selected to assess the object-level method, specifically focusing on footprint semantic matching and quality evaluation based on reference data. For validation, a validation area containing around 5,000 buildings was randomly chosen (Figure 1d).

To assess grid-level completeness, we employed proxy
data, including Esri’s land cover data (Figure 1a) and WorldPop population data (Figure 1b). Building footprint results (Figure 1e) were produced employing a super-resolution-based instance segmentation method (Chen et al., 2023). The reference data (Figure 1f) originated from the 1/2500 scale base map information officially provided by the Geospatial Information Authority of Japan (GSI).

3.2 Grid-level completeness assessment with proxy data

The completeness of extracted footprints was evaluated using a grid-based classification method inspired by Zhou et al. (2022b). This approach involves comparing extracted footprints with proxy data, specifically population count and built area in the land cover data. Initially, we overlaid the two datasets to align extraction results and proxy data. The count of footprints in each basic grid determined subsequent classification. Grid cells were then categorized into four groups based on population count, built-up area, and footprint count.

Type I: No footprints were identified ($N_{fe} = 0$) and had neither a population nor a built area ($N_p = 0$ and $A_b = 0$). (e.g., A1 in Figure 2).

Type II: No footprints were identified but had evidence of a population or the presence of built areas ($N_p \geq 1$ or $A_b > 0$). (e.g., C1 in Figure 2).

Type III: Footprints were identified ($N_{fe} \geq 1$) but had neither a population nor a built area. (e.g., A2 in Figure 2).

Type IV: Footprints were identified and had a population or built area. (e.g., B1 in Figure 2).

where $N_{fe}$ represents the number of extracted footprints and $N_p$ and $A_b$ denote the population and built area in the grid. We assumed that the presence of a building signified either a population or built area.

With the provided classification method, determining the presence of buildings within a grid becomes feasible. Consequently, we adapted the completeness index introduced by Zhou et al. (2022b) to compute the proportion of grid cells containing accurately extracted building footprints in the alternative data (1).

$$C_e = \frac{N_{Type/II} + N_{Type/IV}}{N_{Type/II} + N_{Type/IV}} \times 100\%$$  \hspace{1cm} (1)

where the completeness index $C_e$ represents the footprint completeness assessment factor for the target area. $N_{Type/II}$ and $N_{Type/IV}$ represent the numbers of grids classified as categories II and IV.

To evaluate the accuracy of the grid classification method and the completeness index, we compared the estimated assessment results obtained from reference data. The extracted footprint results were overlaid onto the reference building data by employing the same grid size and layout as those described earlier. Similarly, the grid was divided into four categories. Employing a confusion matrix, we quantitatively evaluated accuracy by comparing classification outcomes using both proxy and reference data. Moreover, to assess the validity of the completeness index, reference index $C_r$ was derived from the reference data. The correlation between the estimated and reference completeness index values for different regions was analyzed through linear regression based on (2). Here, $R^2$ represents the correlation coefficient and $\overline{C_e}$ represents the average of the reference completeness index for each region.

$$R^2 = 1 - \frac{\sum(C_r - C_e)^2}{\sum(C_r - \overline{C_e})^2}$$  \hspace{1cm} (2)

3.3 Object-level matching and quality evaluation
To effectively address geometric inconsistencies and achieve a balance between accuracy and efficiency in large-scale geometric matching, we propose an optimized TWAO method. As illustrated in Figure 3, the optimized method incorporates a secondary matching strategy, where the first matching identifies possible relationships, and the secondary matching refines these matches by superimposing the polygon centroids of each pair. The first matching step employs two metrics, the overlap rate \( R_o \) and centroids distance. We set the thresholds corresponding to a 90% confidence interval of the geometric shifts of buildings in reference data as the criteria. The overlap rate is measured using (3), defined as the intersection area of the two footprint polygons \( A_{overlap} \) divided by the area of the smaller polygon in matching pairs \( A_f \) and \( A_r \) (Fan et al., 2014). If no overlapping footprints were found, other footprints within a specific range of the building centroid were searched, considering the impact of the geometric offset.

\[
R_o = \frac{A_{overlap}}{\min(A_f, A_r)} \tag{3}
\]

After that, the reference buildings in all matched pairs were shifted such that their centroids were aligned with the predicted footprints, to consider the offset of the reference data. Then, the IoU is employed to perform secondary matching (4), where \( A_{overlap} \) and \( A_{union} \) denote the intersection and concatenated areas of the predicted and reference footprints in the first matching.

\[
IoU = \frac{A_{overlap}}{A_{union}} \tag{4}
\]

To evaluate matching result in object-wise, precision, recall, and F-value were selected, as illustrated in (5). These metrics were calculated by quantitatively analyzing the relationship among three categories: TP, false positive (FP), and false negative (FN). Moreover, further analysis was conducted to evaluate the semantic accuracy. Six possible semantic relationships may exist.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{5}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F1 = \frac{2 \times \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

A) 1:1 relation: The predicted footprint corresponded to a single building in the reference data.

B) 1:n relation: The predicted footprint matched multiple buildings in the reference data.

C) 1:0 relation: The predicted footprint did not match any building in the reference data.

D) 0:1 relation: The reference building did not match any footprint in the predicted result.

E) n:1 relation: The reference building matched multiple footprints in the predicted result.

F) n:m relation: Multiple reference buildings and footprints in the predicted results matched.

Figure 4 graphically depicts the assessment of diverse semantic relationships. In datasets lacking correspondence between footprints, 1:0 and 0:1 relationships could be identified. When predicted footprints matched only one reference building, we encountered 1:1 and 1:n scenarios. In such cases, the match of the corresponding reference building was examined in reverse. If the corresponding reference building had a single match, a 1:1 relationship was indicated. Conversely, multiple matches signify a 1:n relationship. Similarly, when the predicted footprint corresponded to multiple reference buildings, n:1 and n:m
relationships were established.

Figure 4 Flow chart and schematic diagram of semantic relationship between the extraction results and reference building.

4. Results

4.1 Grid-level completeness assessment

4.1.1 Grid classification and completeness results

Figure 5 illustrates different grid scenarios: Type I represents the background with no buildings; Type II indicates under-extraction; Type III signifies over-extraction; and Type IV denotes accurate building extraction at the grid level. The completeness assessment results, depicted in Figure 6a, illustrate the grid classification in Hyogo Prefecture. The analysis reveals that a significant majority of the grids fall under type I (61.8%) and type IV (27.7%), with a relatively low proportion of grids mistakenly identified as Type II and Type III at 7.2% and 3.3%. Overall, the extraction results demonstrate an accuracy rate approaching 90%.

Examining the spatial distribution of various grid types, Type IV predominated in the southern coastal area, coinciding with urban agglomerations in Hyogo Prefecture. In contrast, Type III grids were primarily situated in rural and wilderness areas. Type II grids were observed in regions featuring indistinguishable individual buildings and dense vegetation and were also found near topographical variations or built-up areas such as riverbanks. This is often caused by buildings exceeding the grid size (100 m).

Figure 6b presents the completeness of extracted footprint for various cities, represented in 10 bands (0.29–0.96). The histogram reveals that 54 of the 61 cities had completeness >0.6, with approximately half exceeding 0.8, indicating favorable footprint extraction for most urban areas. Spatially, the southern coastal urban agglomeration demonstrates higher completeness, which is consistent with the distribution of Type IV grids.

Figure 5. Typical scenarios representing different classified grids, including (a) Type I, (b) Type II, (c) Type III, and (d) Type IV.

Figure 6. Spatial and frequency distribution results of grid-level completeness assessment in Hyogo prefecture, including (a) grid classification, and (b) estimated city completeness.

4.1.2 Evaluation results

To assess the completeness assessment accuracy, we employed two metrics: grid classification accuracy and correlation between the estimated and reference completeness. Figure 7 presents the confusion matrix and corresponding accuracies for various proxy data in grid classification. Combining population and land use data improved the grid classification accuracy by 1.6% and 2.6% compared to only using only land use and population.
data. Upon analyzing the user’s accuracy (UA) for each grid type, Type I and Type IV exhibited high UAs surpassing 0.9. The lower UA for Type II was due to misclassification due to non-building man-made surfaces (Figure 8a). In regions with a higher building probability like Type IV, this had a lower impact (Figure 8b). Type III presents challenges with reference to buildings in grids that lack population and built areas (Figure 8c). Nevertheless, the UA for Type III improved by 0.4 compared to a population-based approach, mainly due to unpopulated areas (Figure 8d). Incorporating land use data can considerably mitigate this misclassification.

Moreover, Figure 7 shows the relationship between the estimated and reference completeness using different proxy data (a2–c1). Incorporating population and land use data significantly improved the correlation by 0.22 and 0.53. Despite relatively confined result distribution, the linear equation slope was 0.7, indicating proxy data’s tendency to underestimate result completeness (a3–c3 in Figure 7). This is attributed to the confusion matrix for grid classification, leading to an overall underestimation of completeness.

4.2 Object-level quality evaluation

4.2.1 Overall and semantic accuracy results

Manual corrections were applied to the reference buildings within the validation area (Figure 1d) to establish the thresholds. To effectively address geometric errors in the reference buildings and include the majority of buildings in the first matching process, the matching algorithm employed a threshold value corresponding to a confidence level of 0.9. Consequently, the thresholds for the centroid displacement and overlap rate were set to 3 m and 0.4, strictly adhering to the specified requirements (Figure 9). This meticulous threshold selection ensured comprehensive and accurate matching results. The entire matching process for Numazu took 4 min 45 s, resulting in 74,822 TPs among the 92351 buildings. The precision of the extraction results, as listed in Table I was 0.842, whereas the recall was 0.803.
After obtaining the matching results, the semantic relationships between the matched pairs were determined. Figure 10 illustrates the different semantic relationships, and Table II presents the related statistics for Numazu City. The table reveals 74,822 1:1 matching relationships, accounting for 68.7% of the total (Figure 10a). However, 8,402 extracted footprints (1:0) and 8,121 reference buildings (0:1) had no correspondence (Figure 10b and d), because of incorrect extraction and omission by the deep learning model. Furthermore, 17,529 complex semantic relationships existed, including 6,509 reference buildings (6%) with an n:1 relationship (Figure 10e). This suggests that the reference buildings contained more detailed semantic information than the extraction footprints. Additionally, 11,020 relationships, including 10,162 1:n relationships and 858 n:m relationships, had multiple neighboring extracted footprints that matched single or multiple reference buildings, forming a building group (Figure 10c and f).

<table>
<thead>
<tr>
<th>Relation</th>
<th>1:0</th>
<th>1:1</th>
<th>1:n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of matched pairs</td>
<td>8402</td>
<td>74822</td>
<td>10162</td>
</tr>
<tr>
<td>Proportion</td>
<td>7.7%</td>
<td>68.7%</td>
<td>9.3%</td>
</tr>
</tbody>
</table>

4.2.2 Evaluation results
To validate the accuracy of the quality evaluation
results, we assessed the extracted footprints within the validation area and compared the results with those obtained using different matching methods. Table 3 demonstrates that owing to the geometric disparities between the reference building and extraction results, the direct TWAO method exhibited a decrease in precision and recall by 0.080 and 0.114. Conversely, manual matching and the proposed method significantly mitigated the effects of geometric disparities in terms of precision, with marginal decreases of 0.005 and 0.013. The reductions in recall were 0.058 and 0.068, respectively, which were primarily attributed to discrepancies in the acquisition time of the reference data and satellite imagery used for extraction.

Furthermore, we assessed the accuracy of the semantic calculations using the proposed method and measured the precision and recall of various semantic relationships against manually matched semantic relations as the ground truth. As indicated in Table 4, the proposed matching method exhibits an overall matching precision and recall of over 99%. Moreover, for five of the six relationships (excluding n:m), the accuracy exceeded 0.98, validating the effectiveness of the method. The lower accuracy of the n:m relationship was attributed to the limited sample size.

Table 3. Evaluation results of different matching methods.

<table>
<thead>
<tr>
<th>Matching Method</th>
<th>Reference Data</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoU &gt; 0.5</td>
<td>Corrected data</td>
<td>0.872</td>
<td>0.846</td>
<td>0.858</td>
</tr>
<tr>
<td>TWAO</td>
<td>Raw data</td>
<td>0.792</td>
<td>0.720</td>
<td>0.754</td>
</tr>
<tr>
<td>Manual</td>
<td>Raw data</td>
<td>0.867</td>
<td>0.788</td>
<td>0.826</td>
</tr>
<tr>
<td>Proposed method</td>
<td>Raw data</td>
<td>0.859</td>
<td>0.778</td>
<td>0.816</td>
</tr>
</tbody>
</table>

Table 4. Accuracy of different semantic relationships

<table>
<thead>
<tr>
<th>Relation</th>
<th>1:0</th>
<th>1:1</th>
<th>1:n</th>
<th>0:1</th>
<th>n:1</th>
<th>n:m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>1.0</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.84</td>
</tr>
<tr>
<td>recall</td>
<td>0.0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

5. Discussion

5.1 Grid size of completeness assessment

To test the effect of grid size, we assessed the completeness evaluation performance at different spatial resolutions (from 100 m to 1000 m) at 100 m intervals, as shown in Figure 11. First, increasing the grid resolution improves completeness. The average completeness difference between the 1000 m and the 100 m grids was 0.17 (Figure 11a). Second, Figure 11b illustrates the variation in the grid-type proportions with changing grid size. The proportions of the Types III and IV grids gradually increased, whereas those of the Types I and II grids decreased. Moreover, further analysis was performed to examine the effect of grid size on completeness accuracy (Figure 11c). Overall, the R^2 between the estimated and reference completeness values at different spatial resolutions (orange curves) decreased. Larger grid sizes lead to decreased classification accuracy, particularly for misclassified Type III grids, potentially resulting in an underestimation of completeness. So, selecting a 100 m grid size for this study is justified.

Figure 11. Impact of grid size on (a) estimated completeness, (b) different grid type proportions, and (c) completeness correlation and grid classification accuracy.

5.2 Thresholds of building matching

To comprehensively assess the chosen thresholds, we specify ten different sets (ranging from overlap rate 0 to 1 and centroid distance 0 to 10 m), aiming to understand their impact on matching results and semantic accuracy. Figure 12a illustrates the relationship between the overlap rate and precision-recall metrics. As the overlap rate
thresholds increased, TP precision improved from 0.71 to 1, while TP recall decreased from 1 to 0.78. Gradually increasing the overlap rate threshold reduced the number of recognized matching pairs, thus reducing the computation time. Regarding semantic accuracy, the overall precision and recall of the matching process exhibited varying patterns, as illustrated by b1 in Figure 12. Setting a small overlap rate threshold considers even minor overlaps as matches, leading to the misidentification of complex relationships (1:n, n:1, and n:m) and overlooking a relatively large number of 1:0, 1:1, and 0:1 relationships, resulting in lower matching recall and precision. Appropriate overlap rate threshold selection involves a tradeoff between algorithm accuracy and efficiency. Figure 12 suggests a threshold interval of 0.3–0.5 for optimal results and efficiency with higher semantic accuracy and minimized first matches. In contrast, the centroid distance minimally affected the TP determination accuracy, influencing the efficiency by approximately 1%, as shown in Figure 12 (a2). Furthermore, semantic accuracy declined slightly with increasing thresholds because of mistaken matches from the 0:1 to 1:1 relations.

Figure 12. Influence of varying overlapping rate and centroid distance thresholds on (a) accuracy of true positives (TP) and number of matching pairs, and (b) matching accuracy of semantic relationships.

6. Conclusion

In this study, we proposed a two-level approach to evaluate large-scale building extraction results, comprising grid- and object-level assessments. At the first level, we classified the basic grid for each 100 m area using population data from the WorldPop dataset and built area information from the Esri land cover dataset as proxy data. This quantitative analysis helps assess extraction completeness when reference data are unavailable. At the second level, we enhanced the TWAO method by employing the overlap rate and centroid distance as matching metrics and introducing a two-step matching process. This approach effectively mitigated geometric errors and achieved extraction accuracy and semantic precision by meticulously comparing the extracted footprints with reference building data. The second level utilizes available reference data and provides a more detailed evaluation. To validate the effectiveness of the proposed method, we selected Hyogo Prefecture and Numazu City in Japan as the experimental areas. In Hyogo Prefecture, the results showed a 2.6% improvement in grid classification accuracy, and the R² of the estimated and reference completeness increased by 0.53 compared with single proxy data. In Numazu City, the quality evaluation of our method closely aligned with manually derived true values, outperforming the results obtained directly using the TWAO method. Furthermore, the improved matching method achieved a matching precision and recall of >0.99 in semantic relationship recognition, while maintaining high computational efficiency (4 min and 45 s for 92,351 buildings). In summary, our approach effectively assessed the large-scale building extraction results and interpreted the semantic relationships between the extraction results and actual buildings. This study provides crucial support and serves as a valuable reference for decision-making in urban planning and disaster risk assessment. In the future, we aim to enhance the accuracy of grid classification and address the complexities of semantic relationship matching. In addition, we aim to incorporate other dimensions, such as shape similarity, to offer a more
comprehensive evaluation of the results. Moreover, we plan to conduct experiments in diverse research regions, particularly in developing countries, to assess the applicability of this method on a global scale.

Reference


