Developing a Method for Estimating the Distribution of Detached Houses Using Open Data: Toward the Construction of Open Building-Level Spatial Database

Kai Saito[1], Yuki Akiyama[2]

[1]Tokyo City University, Setagaya-ku, Tokyo, Japan – g2481619@tcu.ac.jp
[2]Tokyo City University, Setagaya-ku, Tokyo, Japan – akiyamay@tcu.ac.jp

Keywords: Building Database, Open Data, Machine Learning, Detached Houses, Building Attributes

Abstract

A detailed map database containing attribute information of individual buildings is highly valuable and expected to be utilized in various fields, including urban planning, energy management, and disaster preparedness. However, obtaining such detailed map databases are significant difficulty, because of privacy concerns and their high cost. To address this issue, this research aims to construct an open building-level spatial database as the goal. In this study, as a first step toward achieving this objective, we developed a method to classify buildings into detached houses and other types of buildings by utilizing Foundation Geospatial Data and information derived from open data. First, we assigned explanatory variables to each building in the foundation geospatial data for Nagaoka City, Niigata Prefecture, and created training data using PLATEAU data as the ground truth. Based on this dataset, we developed a machine learning model to classify each building as either detached or other types of buildings. Furthermore, we extrapolated the machine learning model to Sanjo City, Niigata Prefecture. We selected buildings in a way that aligns with the number of detached households reported in the national census at the subregion level and identified these as detached houses. Finally, validation of the extrapolated results showed that the mean absolute error (MAE) at the subregion level was approximately 9 buildings, demonstrating that the model successfully reproduced the spatial distribution of detached houses and other types of buildings.

1. Introduction

In recent years, the need for policy development based on a detailed understanding of urban spaces and their dynamics has become increasingly important. This trend is driven by the global push for cities that are sustainable, disaster-resilient, and resource-efficient. In this context, a detailed map database containing attribute information of individual buildings (hereafter referred to in this paper as the Building Attribute Map (BAM)) is gaining attention as essential infrastructure. They are expected to support a wide range of applications, such as urban planning, energy supply and demand optimization, disaster risk assessment, future population projections, and residential environment analysis. For example, Accurate identification of building usage plays a vital role in optimizing the spatial distribution of urban functions and analysing urban metabolism and resource efficiency, thereby contributing to the development of sustainable cities (Ivanović et al., 2020). In the energy sector, understanding usage patterns by building type has been shown to improve the accuracy of demand forecasting and support the optimization of energy-saving measures (He et al., 2024). Information on building attributes such as structure and year of construction also plays a critical role in estimating damage and assessing impacts during natural disasters such as earthquakes and floods (Rajapaksha et al., 2024). In Japan in particular, seismic performance is known to differ significantly depending on whether a building was constructed before or after the 1981 revision of the Building Standards Act (the introduction of the new seismic design standards). Therefore, identifying the year of construction is effective for improving the accuracy of damage estimation (Takeda et al., 2023). Furthermore, accurately identifying the location of each household and the number of residents is considered useful for understanding the spatial distribution of "shopping-vulnerable" populations and estimating the potential human impact of tsunamis (Akiyama et al., 2013). These examples highlight the critical importance of BAM as foundational infrastructure for urban modeling and policy decision-making. Such data are indispensable for advancing smart city initiatives and promoting digital transformation (DX) in urban governance.

As described above, BAM has the potential to be utilized in various fields. Such data is collected and maintained by local governments, national agencies, and private companies. However, opportunities for these datasets to be made publicly available as open data remain extremely limited. In Japan, BAM is collected and aggregated through the "Basic Surveys Concerning City Planning"(BSCCP) conducted by local governments, and it is regarded as highly valuable foundational data for urban planning (Ministry of Land, Infrastructure, Transport and Tourism of Japan (MLIT), 2023). Nevertheless, due to concerns over personal information protection, public access to this data is restricted. Even when the data is made available as open data, differences in data formats across municipalities pose technical challenges to its utilization. In addition, MLIT is publishing 3D urban models through the "PLATEAU" project. This data utilizes information from BSCCP as building attributes, allowing for a detailed understanding of building characteristics. As a result, it is expected to be useful in applications such as urban design and disaster prevention (MLIT, 2023). However, the considerable time and resources required for data development have meant that coverage has expanded only gradually, on a municipality-by-municipality basis. As of 2025, only 258 cities have made such data publicly available. Although the project aims to cover 500 cities by fiscal year 2027, it does not plan to achieve nationwide coverage. As a result, disparities in data availability across regions remain a significant issue. Meanwhile, BAM developed by private companies also exist. However, these datasets are often expensive, posing a substantial financial barrier to their adoption in research and public administration. Furthermore, many of these data are based on field surveys and are updated infrequently—typically every three to five years—leading to concerns over their timeliness.

As illustrated above, access to and utilization of BAM remains challenging. This limitation continues to hinder broader

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W7-2025-105-2025 | © Author(s) 2025. CC BY 4.0 License.

application across various domains. To address this issue, the development of a detailed and openly accessible BAM holds significant potential as a foundational resource for evidence-based public policy. As a first step toward this goal, this study focuses on building use classification. We propose a classification approach that integrates existing open spatial data with machine learning techniques. This approach aims to contribute to the construction of a high-value urban data infrastructure, with potential applications in municipal smart city initiatives, disaster management, and energy policy.

1.1 Literature Review

Several studies have attempted to estimate building use by utilizing open data. For example, Daniel et al. (2023) developed a method to classify buildings as residential or non-residential using only building shapes from OpenStreetMap (OSM). Fonte et al. (2018) estimated building use in a selected area of Milan, Italy, by utilizing building use and point-of-interest (POI) data from OSM, Facebook, and Foursquare. Fill et al. (2024) developed a highly accurate method for classifying building use not only by utilizing building shapes, land use, and urbanization levels, but also by incorporating spatial relationships into a graph structure. In their approach, buildings were treated as nodes, and distances to their nearest buildings were used as edges to represent spatial relationships. By using this graph structure in a graph neural network (GNN) classifier, high classification accuracy was achieved. However, a common issue in these existing studies is that they use OSM as the ground truth for building use classification. The reliability and quality of OSM labels have not been sufficiently examined or corrected, raising concerns about the validity of the training data.

In another study, Droin et al. (2020) developed a semantic segmentation model (FCN-VGG19) that assigns building use labels to each pixel using aerial imagery and building masks. This enabled building-level use estimation based on pixel-wise classification. Although this study achieved promising results, it has the challenge that collecting source data such as aerial imagery involves considerable time and financial resources.

As described above, several studies have been conducted on building use estimation using open data, but such approaches have rarely been applied in the context of Japan. In Japan, several types of open data are available with high reliability and accessibility, such as the Fundamental Geospatial Data (FGD), which includes nationwide building polygon information, and the national census, which provides insights into regional demographic trends. By combining these datasets with methodologies proposed in previous studies, more accurate building use classification can be expected.

1.2 Objective

The objective of this study is to clarify the distribution of detached houses, which represent the most prevalent building use category in Japan. This serves as a first step toward estimating all types of building use. Specifically, by utilizing publicly available building polygon data and statistical information, this study aims to develop a method for classifying detached houses and other buildings (hereinafter referred to as "non-detached houses") and to understand the spatial distribution of detached houses. The category of "non-detached houses" includes not only apartment buildings but also various other types of buildings such as offices, factories, and schools.

This binary classification approach was intentionally adopted as the first step of a hierarchical strategy, primarily due to the class imbalance within our dataset. As shown in Figure 1, detached houses constitute 76.2% of all instances, making them the overwhelmingly dominant class. In contrast, all other categories are significant minorities (e.g., Public at 5.9%, Apartments at 4.5%, and Commercial at 2.3%). A direct multi-class classification on such an imbalanced dataset would likely result in a model that is heavily biased towards the majority class (detached houses), leading to poor predictive performance for the other building types. Therefore, to build a robust model, we prioritized accurately separating the largest category first. This methodological choice establishes a solid foundation for more detailed, classification of the non-detached category in subsequent research stages.

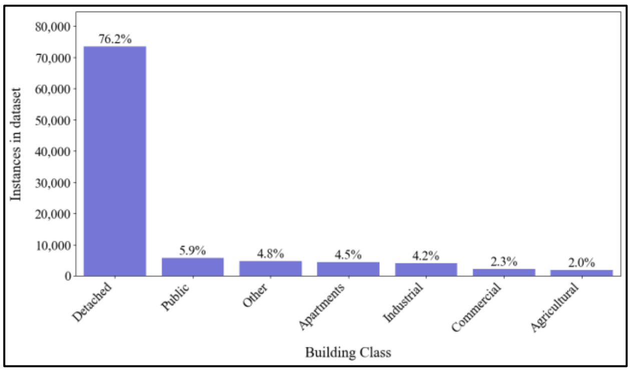


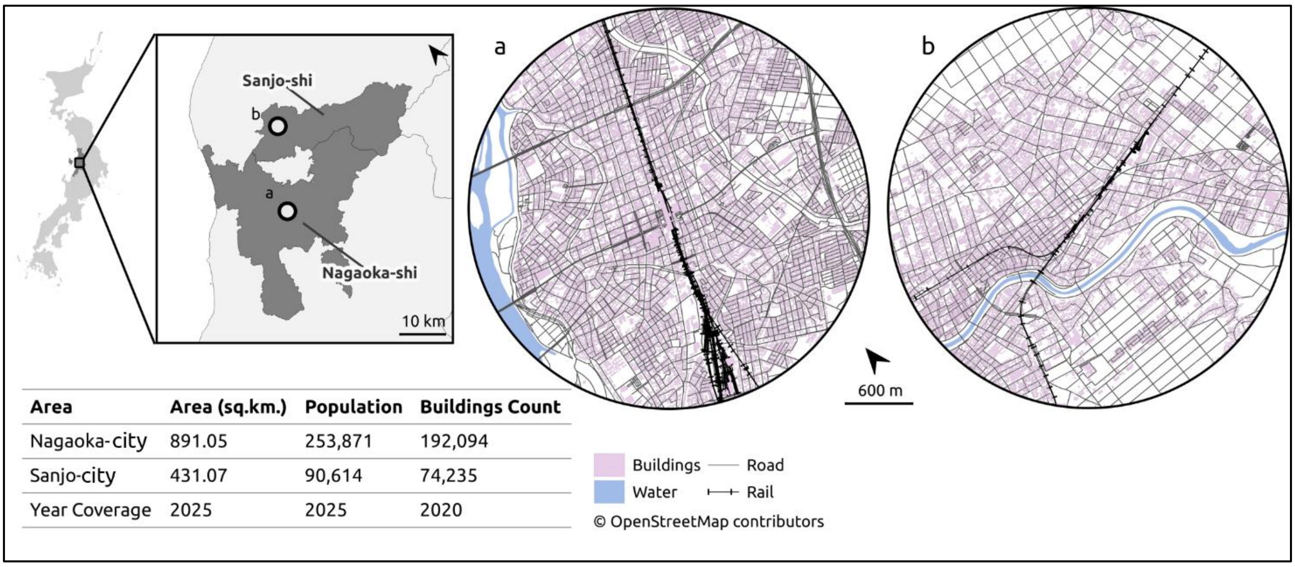
Figure 1. Distribution of building classes

2. Flow of Study

In this study, first, we assigned features to each building in the publicly available building polygon dataset known as the FGD. These features included information derived from the building polygon data, such as building shape and size; demographic information obtained from statistical data; and the distance to POI obtained from OSM. Next, using building use included in the PLATEAU dataset, we labeled each building as either a detached house or a non-detached house and created a training dataset. Based on this dataset, we constructed a machine learning model to classify each building as either a detached house or a non-detached house.Furthermore, we extrapolated the machine learning model to other regions and selected buildings that matched the number of detached households from the national census at the cho-cho-aza (subregion unit in Japan) level, estimating these buildings to be detached houses. Finally, we conducted a validation of the extrapolation results.

2.1 Target Area

The target area of this study is shown in Figure 2. For the creation of the training data used in constructing the classification model, we targeted the entire area of Nagaoka City in Niigata Prefecture, Japan. Nagaoka City is the second-largest city in Niigata Prefecture, with a concentration of public and commercial facilities in its central urban area, while most of the city consists of low-density residential zones. In addition, its suburban areas include rural and mountainous regions, resulting in a diverse land use structure where urban and rural elements coexist. Given this mixture of urban and non-urban characteristics and the diverse



Area	Area (sq.km.)	Population	Buildings Count
Nagaoka-city	891.05	253,871	192,094
Sanjo-city	431.07	90,614	74,235
Year Coverage	2025	2025	2020

Figure 2. Overview of the target area: Nagaoka city and Sanjo city in Niigata Prefecture

distribution of building uses, Nagaoka City is considered well-suited for the creation of training data.

On the other hand, Sanjo City in Niigata Prefecture was selected as the extrapolation target for the constructed classification model. Sanjo City is adjacent to Nagaoka City and shares a similar regional composition. While it has commonalities with Nagaoka City in terms of urban structure and land use patterns, it also possesses distinct regional characteristics. Therefore, Sanjo City is considered suitable for evaluating the generalization performance and regional adaptability of the classification model.

3. Development of a Classification Model for Detached and Non-detached House

3.1 Preparation of Feature Data

The features used in this study consist of 32 variables necessary for classifying buildings as either detached or non-detached houses. These features can be broadly categorized into two groups: geometric characteristics and surrounding environmental characteristics. Geometric characteristics include features such as Area, Perimeter, and Rectangularity. Surrounding environmental characteristics include demographic information such as the proportion of the population by age group, length of residence, and income level within the cho-cho-aza to which each building belongs. These surrounding environmental features were created using data from the 2018 Housing and Land Survey, aggregated at the municipal level, and the 2020 Population Census, aggregated at the cho-cho-aza level. In addition, surrounding environmental characteristics also include features derived from POI data available in OSM. A full list of the features used is provided in Table 1.

3.2 Preparation of Training Data

In this study, we used FGD (2020) as the base building polygon dataset. However, this dataset does not contain any attributes related to building use. To classify buildings as detached or non-detached houses using machine learning, it is necessary to create accurate training data that includes the true building use for each building. Therefore, we spatially joined building use from the PLATEAU dataset—which is based on BSCCP—with the building polygons. Each building was then labeled as either a detached house or a non-detached house, thus completing the preparation of the training dataset.

Category	Feature	Description
Geometric characteristics	Area	Area of the building
	Perimeter	Perimeter of the building
	Rectangularity	Rectangle of the building
Surrounding environmental characteristics	Age categories (Three categories)	Ratio of young population (<18), working age population (18–65), and aging population (>65)
	Period of residence categories (Six categories)	Ratio of each section in which they live
	Family income categories (Nine categories)	Ratio of households in each income group
	Percentage of building use (three categories)	Proportion by building use type (Detached houses, Row houses, Apartment buildings)
	Percentage of structures by construction method (Three categories)	Percentage of each structure in each way of building a house
	Use district	Dummy variable for use district
	Points of interest (POI)	Percentage of each structure in each way of building a house

Table 1. List of input features and their descriptions

3.3 Development of a Classification Model

Using the training data constructed as described above, we developed a classification model for detached and non-detached houses. The model adopted for this task was eXtreme Gradient

Boosting (hereafter referred to as "XGBoost"), chosen for three primary reasons. First, the training data used in this study contains some missing values. XGBoost (Chen et al., 2016) is capable of learning directly from features with missing values, which allows for the construction of a robust model for our dataset. Second, its training speed is extremely fast, significantly reducing the computational cost required for model building. Third, as a tree-based method, XGBoost does not require data standardization. This is because it creates splits based on the relative order of feature values, making it insensitive to differences in their scale or variance. This characteristic, in turn, allows for a much more efficient data preprocessing workflow. For these reasons, we implemented XGBoost as the classification model for this study.

XGBoost is an ensemble method belonging to the "gradient boosting" family. Boosting, in general, is an approach that sequentially trains multiple "weak learners" and combines their results to construct a strong learning model, thereby enhancing predictive performance. A "weak learner" in this context refers to a model that, on its own, does not achieve high performance in classification or regression tasks. Gradient boosting, specifically, considers the error between the predicted output of a weak learner and the true value, and then uses the subsequent learner to correct this error. By iteratively repeating this process, the method progressively reduces the error and builds a robust learning model.XGBoost is a decision-tree-based algorithm, specifically utilizing regression trees. This applies to classification tasks as well, where the final predicted values are converted into prediction probabilities. The following outlines the primary operational principles of the XGBoost algorithm. Initially, a single decision tree is trained based on the given data. The error (or residual) between the predictions made by this decision tree and the actual true values is then calculated. Subsequently, this error becomes the target for the next decision tree to learn. This process is repeated for a specified number of decision trees, bringing the model's overall prediction closer to the target variable. The model's predicted value here is the sum of the weights of the leaves to which the predicted data belongs in each decision tree. Specifically, XGBoost aims to minimize the loss function as described in equation (1).

$$\mathcal{L}(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \qquad (1)$$

where $l\left(\hat{y}_i, y_i\right)$ = error between predicted and true values
$\gamma T + \frac{1}{2}\lambda\|w\|^2$ = regularization term
γ, λ = parameters
T = number of leaves in the decision tree
ω = weights on each leaf

Ultimately, thousands of decision trees are constructed, and by summing the values calculated from each decision tree, the final estimated vacant house probability can be obtained. By adopting such a structure, the model can effectively fit unknown data and return predictive values.

After training the model, it predicts the probability (pi) that each building is a detached house. A building is then classified as a detached house if its predicted probability (pi) exceeds a threshold of 0.5. Logarithmic Loss (logloss) was employed as the evaluation metric for the model's performance. The logloss, defined by the following equation (2), quantifies the distance between the predicted probabilities and the true labels (yi ∈ {0,1}), with N representing the number of data points. A smaller logloss value indicates a superior model performance.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p_i) + (1 - y_i)\log(1 - p_i) \qquad (2)$$

where N = the number of data points
y_i = true values
p_i = probability of being a detached house

For model construction , the entire dataset was first split into training/validation (70%) and testing (30%) subsets. To optimize hyperparameters and enhance generalization performance, we applied Stratified K-Fold Cross-Validation to the training/validation subset. In this approach, the data were divided into three folds (K = 3), with the class distribution (detached / non-detached) maintained consistently across each fold to mirror the overall distribution. In each iteration, one fold was used for validation while the remaining two were used for training. This process was repeated three times to ensure a consistent and reliable evaluation of the model.

Furthermore, the model was optimized to maximize the F1-score, and key hyperparameters were tuned accordingly. Specifically, we optimized the maximum depth of the decision trees, the learning rate, the subsampling ratio for data instances, and the column subsampling ratio for features. Table 2 summarizes the types and search ranges of the hyperparameters considered, along with the optimal values identified in this study. Of note is the maximum tree depth, for which the optimal value was found at the upper boundary of our search range (10). While this might suggest that a deeper tree could yield better performance, we intentionally set the upper limit at 10 to manage computational costs and mitigate the risk of overfitting. Trees with a depth greater than 10 are prone to overfitting the training data, which can impair the model's generalization performance on unseen data. Therefore, after considering the trade-off between generalization performance and model complexity, we determined this range to be a practical and appropriate setting for our study's objectives.

Hyperparameter	**Search range**	**Optimized value**
Maximum tree depth	3 - 10	10
Learning rate	0.01 - 0.5 (log scale)	0.0383
Subsample ratio	0.5 - 1.0	0.868
Column subsampling per tree	0.5 - 1.0	0.623

Table 2. Summary of tuned hyperparameters used in this study

3.4 Classification Results

The estimation results for the test data are shown in Table 3. The overall accuracy was approximately 79%, and the F-score was 0.804, indicating that the classification model developed in this study can perform high-accuracy classification. On the other hand, some buildings were found to be misclassified. One possible reason for this is the use of geometric characteristics as features. One possible reason for this is the use of geometric characteristics as features. As a result, it is possible that detached houses with large areas and complex shapes were misclassified as non-detached buildings, while non-detached buildings with small areas and simple shapes were mistakenly classified as detached houses. Among the misclassified cases, the most frequent error involved non-detached buildings being incorrectly predicted as detached houses. This may be attributed to the use of aggregated data at the cho-cho-aza level as surrounding environmental features, which might not have accurately reflected the specific characteristics of individual buildings within cho-cho-aza. aggregated data at the cho-cho-aza level as surrounding environmental features, which might not have accurately reflected the specific characteristics of individual buildings within cho-cho-aza.

Figure 3 shows the classification results for the area around Nagaoka Station in Nagaoka City. As can be seen from Figure 2, many buildings located in the central urban area around Nagaoka Station were classified as non-detached houses, while detached houses were found to be distributed in the surrounding areas. As described in Section 2.1, public and commercial facilities are concentrated near Nagaoka Station, and much of the city consists of residential zones. Therefore, the classification results appropriately reflect the actual building use in the area, demonstrating the model's high classification accuracy.

Figure 4 visualizes the model's classification results for the area around Nagaoka Station, with each building color-coded by its confusion matrix category. The map shows that correctly classified buildings—True Positives (TP) and True Negatives (TN)—account for the majority of all instances. However, it is also apparent that a non-negligible number of False Positives (FP) are present. A closer inspection reveals that many of these FP buildings, which were misclassified as detached houses, tend to have areas and shapes similar to those of actual detached houses.

Figure 5 shows the top 10 features that contributed to the classification of building types. While it is difficult to fully explain the causal relationships behind how these features affected the model’s classification results, it is possible to infer their roles to some extent by considering the meaning and tendencies of each feature. However, it is possible to infer their roles to some extent by considering the meaning and tendencies of each feature. In this study, we used Shapley values (SHAP) to visualize how each feature contributed to the classification decisions (Scott M. Lundberg, 2017). As shown in Figure 5, both geometric characteristics and surrounding environmental characteristics play an important role in the classification accuracy. Regarding geometric characteristics, there was a tendency for buildings with larger values for *“Area”*, “*Perimeter”*, and “*Rectangularity”* to be classified as non-detached houses. This suggests that buildings with more complex shapes are more likely to be identified as non-detached houses. As for surrounding environmental characteristics, there was a tendency for buildings to be classified as non-detached houses in cho-cho-aza where the proportion of long-term residents was higher—particularly those falling under the period of residence categories "Resident Since Birth" and "Resident for More Than 20 Years". In addition, there was a tendency for buildings to be classified as non-detached houses in areas where the values were lower for the categories "Detached House" (Percentage of Building Use) and Percentage of Structures by Construction Method "Standard Building"(Percentage of structures by construction method). As for the Age categories, there was a tendency for buildings to be classified as detached houses in areas with a higher proportion of the “Young Population” and also in areas with a lower proportion of the “Aging Population.” Based on these results, it can be concluded that both the physical shape information at the building level and the statistical data available at the cho-cho-aza level function effectively in the classification of building use.

Test data		Predicted	
		Non-detached	Detached
True	Non-detached	16,218	7,232
	Detached	2,379	19,677
Accuracy			0.789
Precision			0.731
Recall			0.891
F-score			0.804

Table 3. Confusion matrix and performance metrics on our classification model

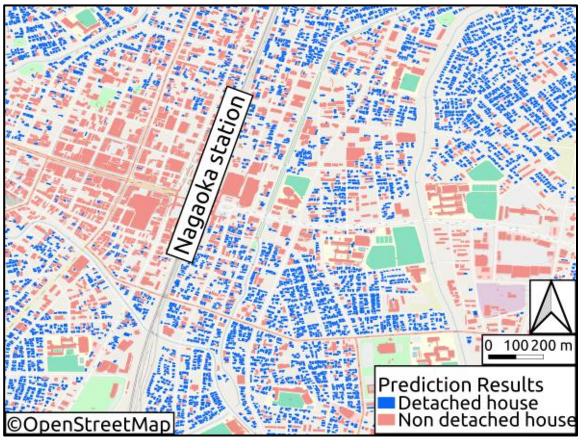


Figure 3. Estimation results in Nagaoka city, Niigata prefecture (example of Nagaoka Station area)

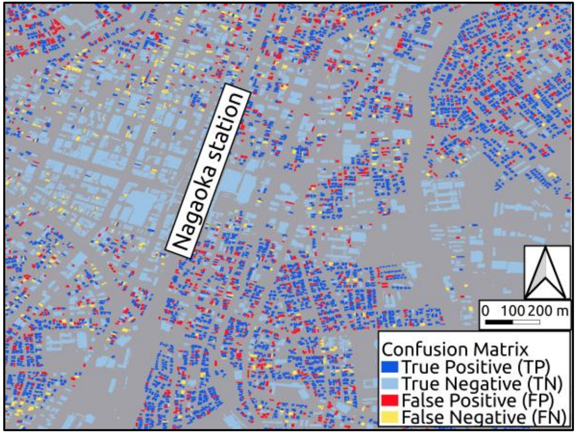


Figure 4. A spatial distribution of the confusion matrix in Nagaoka city, Niigata prefecture
(example of Nagaoka Station area)

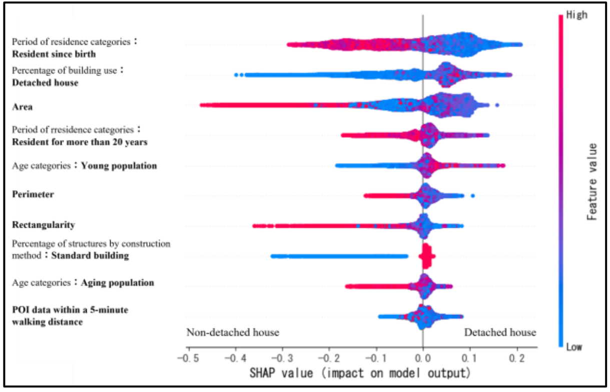


Figure 5. Top 10 feature contributions to the detached/non-detached classification model, based on SHAP summary plot

https://doi.org/10.5194/isprs-annals-X-4-W7-2025-105-2025 | © Author(s) 2025. CC BY 4.0 License.

4. Extrapolation

4.1 Extrapolation to Other Regions and Validation of Extrapolation Results

In this study, we extrapolated the developed classification model to all buildings within Sanjo City, Niigata Prefecture, in order to estimate the probability that each structure is a detached house. At the initial stage, a threshold-based approach was adopted, whereby buildings with a model-derived probability of 0.5 or higher were classified as detached houses. When the extrapolated results were aggregated at the cho-cho-aza level, this threshold-based method led to biases in the estimates: in some areas, the number of detached houses was overestimated, while in others, none were identified at all. As shown in Figure 6, this issue became evident through validation against the digital residential map (Zmap TOWN II: 2020), resulting in a mean absolute error (MAE) of approximately 42 and a mean absolute percentage error (MAPE) of 336.7%, indicating considerable discrepancies and regional variability.

This estimation error is primarily attributed to the use of aggregated statistical data at the cho-cho-aza level as model features, which introduced a regional bias into the predicted probabilities. In other words, aggregated data has the limitation of "averaging out" the diversity and micro-locational environment of individual buildings, thereby failing to capture their unique characteristics. Indeed, our results confirmed this phenomenon, showing that even for buildings with similar geometric shapes, their classification probabilities differed significantly depending on their location. To address this, we revised the extrapolation method by aligning the estimated number of detached houses with the figures from the 2020 Population Census at the cho-cho-aza level. Specifically, within each area, buildings with higher probabilities were selected in descending order until the census-based count was matched. As illustrated in Figure 7, this adjustment improved estimation accuracy, especially in previously over- or under-estimated areas, reducing the MAE to approximately 8.8 and the MAPE to 12.5%. It should be noted, however, that the Population Census is based on household counts, and may not fully account for dwellings with multiple households, such as two-family homes. Taking this into consideration, the extrapolated results shown in Figure 8 indicate a slight overestimation of detached houses compared to the digital residential map in many areas. Nonetheless, this suggests that the model accurately reflects actual residential patterns and performs well in terms of classification accuracy.

To further assess the validity of the extrapolation results, the absolute error between the extrapolated number of detached houses and the digital residential map was visualized for each cho-cho-aza. Figure 8 shows the distribution of absolute errors based on the threshold-based method, revealing substantial variation in errors across different regions. In contrast, Figure 9 presents the distribution of absolute errors using the census-based method, where the errors were reduced to fewer than 20 buildings in most areas. These results demonstrate that the initial threshold-based method caused large regional errors and spatial bias, whereas the adjustment using census data effectively mitigated these issues. Based on this process, Figure 10 shows the extrapolated results around Higashi-Sanjo Station in Sanjo City.

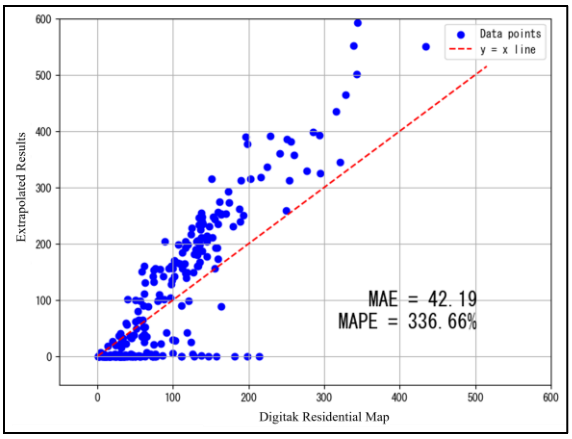


Figure 6. Comparison of extrapolation results and digital residential map (Threshold-based method)

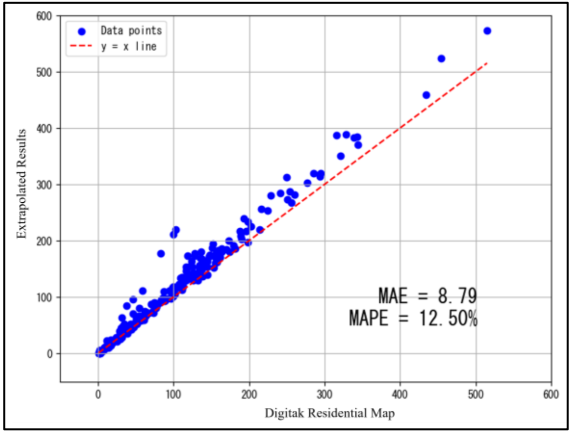


Figure 7. Comparison of extrapolation results and digital residential map (Census-based method)

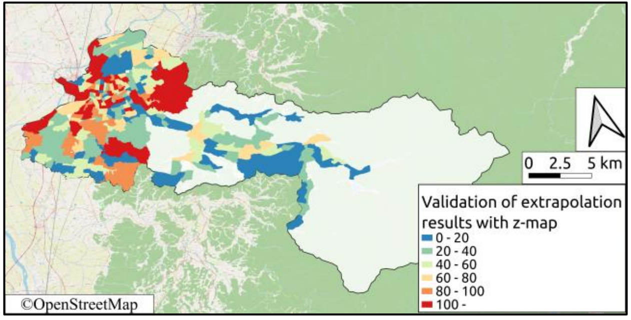


Figure 8. Distribution of absolute errors between extrapolated results and digital residential map (Threshold-based method)

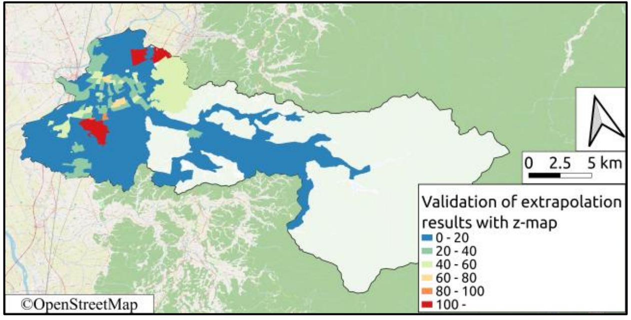


Figure 9. Distribution of absolute errors between extrapolated results and digital residential map (Census-based method)

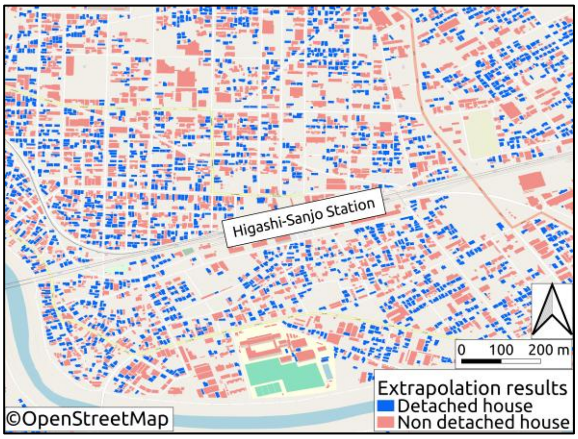


Figure 10. Estimation results in Sanjo City, Niigata Prefecture (example of Higashi Sanjo Station area)

5. Conclusion

In this study, we developed a method for classifying buildings as either detached or non-detached houses by utilizing open statistical data and POI data from OSM, based on the publicly available building polygon data from the FGD. As a result, the spatial distribution of detached houses could be identified with high accuracy. However, the model, trained on data from a single regional city, shows limited generalizability when extrapolated to large metropolitan areas. This issue stems from the fact that the statistical features used are not consistent across different urban scales. To overcome this and enable nationwide application, our future work will involve developing separate models tailored for distinct 'urban,' 'suburban,' and 'rural' typologies. We also aim to refine the classification of non-detached houses by distinguishing specific types such as apartment buildings and commercial facilities. In particular, we are considering applying the graph neural network (GNN) classifier method proposed by Fill et al. (2024), which accounts for spatial relationships among buildings, to develop a use classification model tailored to regional characteristics in Japan.

The ultimate goal of this research is to construct a detailed, nationwide building database. To achieve this goal, a key advantage of our methodology is its ability to build this database efficiently and at a low cost by leveraging open statistical data and building polygons that are available across the country. The construction of this database will involves estimating key building attributes—temporal attributes (e.g., construction year), thematic attributes (e.g., building structure), and spatial attributes (e.g., height and number of floors)—which are then combined to infer the location and number of residents for each household. Through these estimations, we aim to build an open map database that integrates detailed information on individual buildings at a nationwide scale. Such a database is expected to serve as a foundational resource for enhancing analysis and decision-making in a wide range of fields, including urban planning, energy policy, and disaster risk assessment. Beyond academic use, this resource is also expected to contribute to the realization of smart cities by supporting digital transformation (DX) in municipal planning and promoting the visualization and understanding of dynamic and complex urban spaces.

Acknowledgements

This research was the result of the joint research with CSIS, the University of Tokyo (No. 990) and used the following data: Zmap TOWN II data provided by Zenrin CO., LTD.

References

Adams, D.S., Hauser, T., Moehl, J., 2023: Decoding Ethiopian Abodes: Towards Classifying Buildings by Occupancy Type Us ing Footprint Morphology. Proc. 2023 Int. Conf. on Machine Le arning and Applications (ICMLA), 210–217. doi.org/10.1109/IC MLA58977.2023.00037

Akiyama, Y., Takada, H., Shibasaki, R., 2013: Development of Micropopulation Census through Disaggregation of National Population Census. Proc. CUPUM 2013 – 13th International Conference on Computers in Urban Planning and Urban Management, 2–31.

Chen, T., Guestrin, C., 2016: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Droin, A., Wurm, M., Sulzer, W., 2020: Semantic labelling of building types. A comparison of two approaches using Random Forest and Deep Learning. Publ. DGPF, 29, 527–538.

Fill, J., Eichelbeck, M., Ebner, M., 2024: Predicting building typ es and functions at transnational scale. arXiv:2409.09692. doi.or g/10.48550/arXiv.2409.09692

Fonte, C.C., Minghini, M., Antoniou, V., Patriarca, J., See, L., 2018: Classification of Building Function Using Available Sources of VGI. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-4, 209–215. doi.org/10.5194/isprs-archives-XLII-4-209-2018

He, Z., Yao, W., Shao, J., Wang, P., 2024: UB-FineNet: Urban Building Fine-grained Classification Network for Open-access S atellite Images. Preprint, 14 pp. doi.org/10.48550/arXiv.2403.02 132

Ivanović, B., 2020: Multi-functional Land-use Planning as a Regulator of Urban Metabolism: A Conceptual Perspective. SPATIUM, 43, 52–58. doi.org/10.2298/SPAT2043052I

Lundberg, S.M., Lee, S.I., 2017: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst., 30, 4765–4774. doi.org/10.48550/arXiv.1705.07874

Ministry of Land, Infrastructure, Transport and Tourism (MLIT), 2022: Current Status and Challenges of Urban Planning GIS. https://www.mlit.go.jp/toshi/tosiko/content/001510040.pdf , (19 April 2025)

Ministry of Land, Infrastructure, Transport and Tourism (MLIT), 2023: On the Development, Utilization, and Open Data Promoti on of 3D Urban Models (Project PLATEAU). https://www.mlit. go.jp/policy/shingikai/content/001609086.pdf, (19 April 2025)

Rajapaksha, D., Siriwardana, C., Ruparathna, R., Maqsood, T., Setunge, S., Rajapakse, L., De Silva, S., 2024: Systematic Mapping of Global Research on Disaster Damage Estimation for Buildings: A Machine Learning-Aided Study. Buildings, 14(6), 1864. doi.org/10.3390/buildings14061864

Takeda, N., Furuya, T., Akiyama, Y., 2022: Development of Est imation Method for Building Structure using Open Data and Sta tistics. Proc. IGARSS 2022 – IEEE International Geoscience an d Remote Sensing Symposium, 2439–2441. doi.org/10.1109/IG ARSS46834.2022.9883801