Accuracy of vacant housing detection models: An empirical evaluation using municipal and national census datasets

Kanta Sayuda1 | Euijung Hong1 | Yuki Akiyama2 | Hiroki Baba3 | Tomoya Tokudomi4 | Toshihiko Akatani5

1Department of Urban Engineering, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
2Department of Urban and Civil Engineering, Faculty of Architecture and Urban Design, Tokyo City University, Tokyo, Japan
3Center for Southeast Asian Studies/Hakubi Center for Advanced Research, Kyoto University, Kyoto, Japan
4Wakayama Data Utilization Promotion Center, Wakayama, Japan
5Statistics Bureau, Ministry of Internal Affairs and Communications, Tokyo, Japan

Correspondence
Kanta Sayuda, Department of Urban Engineering, Graduate School of Engineering, The University of Tokyo, Hongo7-3-1, Bunkyo Ward, Tokyo 113-8656, Japan.
Email: mono9649.cho56@gmail.com

Abstract
In Japan, the rise in vacant housing has created the need to develop quick, effective, and inexpensive methods to detect the spatial distribution of vacant housing at the municipal level. However, due to incomplete and inaccessible data, the change in the accuracy of the vacant housing detection model must be evaluated while accounting for the limited data. Therefore, this study compares the performance of vacant housing detection models for different data combinations (Basic Resident Register; building registration, water usage, and national census) by considering Wakayama City, Japan, as the case study setting. Three main findings emerged: (1) the contribution of the data to the accuracy varies with the combination of datasets and metrics; (2) even if specific municipal data are unavailable, it is possible to acquire a similar accuracy by combining other data; and (3) the missing value contributes to the vacant housing detection rather than the feature value itself.

1 | INTRODUCTION

An increase in vacant housing, caused by an excess in housing supply compared to the demand (Zabel, 2016), has emerged as a critical global problem (Mallach et al., 2017; Martinez-Fernandez et al., 2016; Schwarz et al., 2010). The decline in housing demand is mainly associated with a decline in population; thus, vacant housing is widespread...
in cities with shrinking populations (Couch & Cocks, 2013; Hollander et al., 2018; Radzimski, 2016). Accordingly, urban shrinkage, as advocated by Oswalt (2006), is especially prevalent in developed countries characterized by a declining population, such as the United States (Schilling & Logan, 2008), Germany (Bontje, 2004), and Japan (Hattori et al., 2017), and is spreading to cities worldwide.

An excessive number of vacant houses causes various problems, such as an increase in crime and a decrease in the vitality of the neighborhood environment (Abe et al., 2014; Baba & Asami, 2017). Accordino and Johnson (2000) found that the increase in vacant houses and vacant lots in American cities has significantly affected its commercial and neighborhood vitality and crime prevention efforts. Similar effects have also been observed in other cities in the United States (Zahirovich-Herbert & Gibler, 2014), East Germany (Bernt, 2019; Radzimski, 2016), and South Korea (Nam et al., 2016). Furthermore, dilapidated and abandoned houses are sometimes known to negatively impact the property value of neighboring houses (Han, 2014).

According to the Housing and Land Survey published by Japan’s Ministry of Internal Affairs and Communications, the vacant housing ratio has been rising continuously since 1958, when the survey began, and achieved its peak at 13.6% in 2018. Technological improvements have generally improved the performance of newly constructed buildings, which lowers the market value of the existing housing stock (Thomsen & Van der Flier, 2011). Therefore, in Japan, once an old house becomes vacant, it is less likely to be preferred by new residents (Akiyama et al., 2019).

In 2015, Japan enacted the Law on Special Measures to Promote Municipal Initiatives on Vacant Houses, which enables local governments to plan countermeasures for vacant houses. Accordingly, each municipality needs to identify the spatial distribution of vacant houses. The primary method of ascertaining the vacant housing is to visit individual buildings and visually judge their appearance. However, on-site surveys require considerable time, money, and labor. In addition, the Housing Bureau of the Ministry of Land, Infrastructure, Transport, and Tourism (MLIT, 2018) emphasized the difficulty of determining the presence of occupants through a visual assessment alone, which makes the survey problematic. Since no practical method has been established for local governments to survey the distribution of vacant houses, there is a need for quick, easy, and inexpensive methods to identify and estimate the spatial distribution of vacant houses in Japan.

To approach these challenges, the use of public data has received much attention (Cheshire et al., 2018). For instance, Akiyama et al. (2019) estimated the vacant housing ratio of Japanese cities using closed municipal data with the average vacant housing ratio for each 250-m square grid. Similarly, Baba et al. (2020) conducted a machine-learning-based estimation of vacant houses for each building unit using various municipal information, such as building and resident register data. However, not all municipalities maintain such datasets. Hence, to detect vacant houses effectively during limited data availability, researchers must first verify the extent to which the accuracy of the data is affected using specific combinations of datasets. In addition, in the black-box machine learning process, determining the value of these datasets in detecting vacant houses is difficult. Therefore, models used to detect vacant houses must be analyzed for their interpretability.

Considering the social needs and the research gap, this study evaluates the accuracy of models for detecting vacant houses using various types of municipal datasets. Although public data openly provide nationwide information on various aspects, the units are coarsely aggregated like municipal or prefectural levels. In contrast, municipal data provide more detailed information at a higher resolution. By examining the extent to which the differences in the original data alter the model’s predictive accuracy, we offer an appropriate combination of municipal datasets under the existing data availability constraints. We also provide insights into the correlation between specific datasets and vacant housing detection, so that which dataset better contribute to the search of vacant housing.

2 | BACKGROUND ON VACANT HOUSING DETECTION

The existing literature shows three trends in vacant housing detection methods: (1) field surveys; (2) data-driven analysis; and (3) machine-learning-based detection. Ishikawa and Akagawa (2014) and Kubo and Mashita (2015) analyzed
the management status of vacant houses using the field survey method. Similarly, Kato et al. (2009) examined the renovation methods and mechanisms of vacant houses, and Sasaki et al. (2010) investigated the willingness of vacant house owners to donate their houses through questionnaires. Although these studies analyze the characteristics of vacant houses from various perspectives, they all relied on field survey methods such as a visual inspection of buildings and interviews with local government representatives. These methods are unsuitable for many municipalities with small budgets and due to the personnel cost involved in conducting continuous investigations of vacant houses in large areas.

Second, researchers have employed various data-driven approaches. The utilization of novel datasets that has not been utilized enables rapid investigation of the spatial distribution of vacant houses in large areas. Bourne (2019) used public data to estimate the extent and value of underutilized domestic real estate. Du et al. (2018) estimated the census-level vacant housing ratio by analyzing nighttime satellite imagery. Kanamori et al. (2015) constructed a vacant housing ratio regression model based on census data on housing stock, construction, and population composition. Such methods help local governments to efficiently understand the spatial distribution of vacant houses. However, their data aggregation unit is too large to build an accurate model. In this respect, some previous studies have utilized closed municipal data to improve the accuracy of detecting vacant houses. For example, Ishikawa et al. (2017) estimated the vacant housing ratio in each city block based on the number of dwelling units and households. Nevertheless, their results were overestimated compared with the field survey results. Akiyama et al. (2019) estimated the vacant housing ratio in Japanese cities using closed municipal data. Although the model performed accurately in small areas, the model still calculated the vacant housing ratio per grid area. Therefore, there is room for improvement in the resolution of vacant housing detection.

Moreover, the use of machine learning methods (Athey, 2017) helps to deal with the use of various types of data—from census to municipal information—maximize the accuracy of the model estimation, and develop evidence-based policies (Howlett, 2009). Recently, Baba et al. (2020) built a vacant housing detection model using machine learning methods with many closed datasets: basic resident registration, building registration, and monthly water usage. The model helped detect vacant houses in each building unit with an accuracy rate of 95.4%, which was sufficiently high for detecting vacant houses. However, this study presents two challenges in operationalizing it in practice. First, researchers may not be able to use all of the data in other municipalities under data availability constraints. Thus, researchers need to verify whether the model can maintain accuracy, even with limited datasets. Furthermore, the accuracy of a machine learning model strongly depends on its objectives and dataset selection (Ularu et al., 2012). As the volume of data will continue to compound in the future, researchers cannot continue to operate with the attitude of "more data usually beats better algorithms." An efficient selection of datasets is crucial for the efficient use of machine learning (Stanula et al., 2018). Additionally, interpreting machine-learning-based models is challenging because of its black-box computational process. The low interpretability of models undermines their reliability in practice because it creates ambiguity in identifying the data that contributed toward determining vacant houses. Recent studies have attempted to increase the transparency of models by visualizing the results (McGovern et al., 2019) and selecting their key features (Chen et al., 2019). Considering these challenges, this study analyses the accuracy of models derived from the different combinations of datasets. While the aim of the study to use data for speedy and accurate detection of vacant houses is similar to several previous studies, our study extends the feasibility of the model to operationalize it in practice by presenting more practical options for data combinations and improve the model's interpretability.

3 | DATA

3.1 | Study area

Using closed datasets provided by the municipality, we considered Wakayama City, Japan, as the case study setting. Wakayama City is the capital of Wakayama Prefecture, Japan, with a total land area of 208.85 km square. Area-wise,
Wakayama City occupies only 4% of the Wakayama Prefecture, whereas its population accounts for approximately 40% of the total population of the Wakayama Prefecture, making it a regional center. The total population of Wakayama City peaked at 401,352 in 1985 and then began to decline. Recently, its population has been decreasing by more than 2000 people every year. As of July 1, 2021, its population had declined to 355,686. The city center is also suffering from population decline and the problem of vacant housing. This pattern is prevalent in many provincial cities in Japan. In this sense, Wakayama City is a representative and suitable research area (see Figure 1). The western part of Wakayama City, along the sea, is a fishing village. The eastern side has mountains and mountain villages. The city center is located along the railroad station.

### 3.2 | Data description

To predict the spatial distribution of vacant houses, we employed four datasets: the Basic Resident Register data, building registration data, and water usage data held by local governments, as well as the national census data from the national government. The vacant house field survey data are obtained from Wakayama city, and are used as a supervised data.

- **Basic resident register data (RR data)**
  
  The basic resident register contains resident details such as name, date of birth, sex, address, and so on. In Japan, residents must have a certificate of residence at the place of their residence by law. In this study, we used the data from Wakayama City’s RR data as of April 2019. These data include features such as the number of household members, age, and date of the moving-in declaration for each address.

- **Building registration data (BR data)**

![Figure 1: Geographical location of Wakayama city](image-url)
BR data include data on various aspects of the building, such as building type, condition, area, and structure. For this study, we used the BR data as of October 2018.

- **Water usage data (WU data)**
  This dataset contains the monthly WU data for each address held by the Waterworks Bureau of Wakayama City. In addition to monthly usage, these data include information on the state of the water contract including its start and end date. In this study, we included water supply information data as of May 2019.

- **Vacancy house field survey data (VH data)**
  In Wakayama City, vacant houses were detected by field surveys. Possible vacant houses were selected based on WU data from the 2012 Wakayama Prefecture Abandoned Buildings Survey and the Wakayama City Dangerous Houses Register (a register that records complaints). In the field surveys, vacant houses were identified through visual confirmation of the building exterior. We used the data from 2016 to 2017 as the supervised data. Figure 2a shows a map of the number of vacant houses in a 500-m grid cell, and Figure 2b shows a map of the ratio of vacant houses in a 500-m grid cell. Areas closer to the city center have a higher number of vacant houses. However, since the city center area is expected to have a higher housing density, the vacant housing ratio should also be examined. Accordingly, Figure 2a shows that the vacant housing ratio is also higher in areas close to the city center, indicating that the city center area is now in decline. In contrast, Figure 2b confirms that the vacant housing ratio is not as high in the suburbs close to the city center, but rises again in the distant suburbs far from the city center.

- **The National Census (NC data)**
  In this study, we used the National Census of Japan as of 2015, which is the latest data available. The NC data include data on administrative and other measures to ascertain the actual conditions of the country’s citizens and households. The census covered each household. However, to protect citizens' personal information, this study used data aggregated in "basic unit districts" of 20–30 households. Therefore, this study handled census data differently from other data included in this study. The data were retrieved from [https://www.stat.go.jp/english/data/kokusei/index.html](https://www.stat.go.jp/english/data/kokusei/index.html)

It is true that we acknowledge the time periods of the data collection differ, but we consider that data alignment issues will not affect the results. We chose datasets to be the closest and most recent to 2017, the year of the supervised data, and they are not strictly at the same time period. Specifically, the BR data is 6 months older, and the NC data are 4 years older than the RR and WU data. However, the change in the features of both BR and NC data should be sluggish, so we assume that household composition and building characteristics are not expected to change significantly within a few years. Nevertheless, the age of the NC data is an issue to be addressed, and it needs to be updated when the new data become available.

### 3.3 Data integration

This study integrated multiple datasets into one based on building units for the following steps. First, we extracted the polygon data with "detached house" category from the building polygon data obtained from Zmap TOWN II by Zenrin Co. Ltd. Second, the RR, BR, and WU data were geocoded and represented as point data. Finally, the respective data points were integrated by a spatial join to the corresponding building polygons (Figure 3).

Each value of the data was assigned to correspond with the building polygons with the closest distance from each of the WU, RR, BR data coordinates. Since the geographical locations of the data are originally indicated by text-based address, we geocoded them using the CSV address matching service provided by Center for Spatial Information Science, the University of Tokyo ([http://newspat.csis.u-tokyo.ac.jp/geocode/](http://newspat.csis.u-tokyo.ac.jp/geocode/)). In contrast, since the VH data already contained the coordinates, the result of the VH survey was applied to the building polygon with overlapping coordinates. Consequently, the data integration work was completed using the unique ID of the building polygon as
Figure 2 (a) Number of vacant houses; and (b) Vacant housing ratio No grid indicates that there is no building in the area.
the critical feature. The resulting data made it possible to create analytical models, estimate vacant houses, and verify the accuracy of the estimation.

Despite following the process as mentioned above, we were unable to fully integrate the different types of data used in this study. This could be due to incomplete addresses registered in the original data and the limited accuracy of the matched addresses. Furthermore, there is a case that VH point data does not overlap with a building such as a U-shaped polygon. Therefore, we assumed that our analysis contained a certain number of coordinates without a perfect match. However, because more than 60,000 samples were geocoded in each dataset, we could obtain sufficient matches for the analysis. Table 1 shows the data integration results of each dataset, and Table 2 shows the attributes of the vacant house database.

### 4 | METHODS

#### 4.1 | XGBoost

In this study, we employed XGBoost, proposed by Chen and Guestrin (2016), to classify vacant houses. XGBoost is a machine learning method based on a gradient boosting decision tree. Its sequential process called gradient boosting can efficiently improve model accuracy.

The applicability of gradient boosting decision tree-based models is generally higher than that of other estimation methods. First, logistic regression and other generalized linear models are easy to interpret because they assume that the number of vacant houses increases (or decreases) monotonically with respect to features such as
| Data source                      | Year | Attribute                        | Data type | Sample |
|---------------------------------|------|----------------------------------|-----------|--------|
| Field survey (VH)               | 2016 | Vacant house dummy               | C         | 1      |
| Zmap TOWN II (LU Data)          | 2020 | Zoning district                   | C         |        |
|                                 |      | Building coverage ratio           | V         | 60%    |
|                                 |      | Designated floor area ratio       | V         | 200%   |
| Basic resident register (RR data) | 2019 | Maximum age in a household       | V         | 53     |
|                                 |      | Minimum age in a household        | V         | 16     |
|                                 |      | Number of residents               | V         | 4      |
| Building registration (BR data) | 2018 | Housing dummy                    | C         | 1      |
|                                 |      | Building structure                | C         | Wooden |
|                                 |      | Building age                      | V         | 6      |
|                                 |      | Gross floor area                  | V         | 101.85 |
|                                 |      | Number of stories                 | V         | 1      |
| Water usage (WU data)           | 2019 | Average water use                 | V         | 263.3  |
|                                 |      | Whether the hydrant is open       | C         | 1      |
|                                 |      | Number of months since the hydrant was closed | V | NA |
| National census (NC data)       | 2015 | Type of household                 | C         | General household |
|                                 |      | Type of residence                 | C         | Detached house |
|                                 |      | Number of stories                 | C         | One or two stories |
|                                 |      | The household's floor number      | C         | First or second floor |
|                                 |      | Type of housing and ownership     | C         | Owned house |
|                                 |      | Number of residents in a household | C       | 4      |
|                                 |      | Type of family members            | C         | Family of parents and children |
|                                 |      | Type of households with residents aged 65 years and above | C | Households with two members (married couples only) |
|                                 |      | Number of children                | V         | 2      |
|                                 |      | Maximum age of the children       | C         | 17     |
|                                 |      | Minimum age of the children       | C         | 16     |
|                                 |      | Number of male residents          | V         | 2      |
|                                 |      | Number of female residents        | V         | 2      |
|                                 |      | Marital status                    | C         | Married |
|                                 |      | Nationality                       | C         | Japanese |
|                                 |      | Period of residence               | C         | 10–20 years |
|                                 |      | Place of residence before 5 years | C         | Another province |
|                                 |      | Work status                       | C         | Labor |
|                                 |      | Commute status                    | C         | Commute to same city |
|                                 |      | Employee status                   | C         | Self-employed |
|                                 |      | Type of industry                  | C         | Transportation industry |

Note: "V" in the "data type" represents numerical data, and "C" represents categorical data. Categorical attributes were converted into dummy variables. Regarding the NC data, each feature is divided into several classes and then aggregated in a basic unit district. Consequently, 179 features were used for the analysis (See Appendix A).
building area and age. However, as these models do not aim to improve the prediction accuracy, which this research mainly focuses on, some studies showed that XGBoost had substantial advantages for prediction accuracy (Mo et al., 2019; Ogunleye, & Wang, 2020). Second, other machine learning methods such as Neural Networks or Random Forest techniques generally guarantee high accuracy. However, the models above cannot handle missing values and need imputation preprocessing. XGBoost, on the other hand, is able to treat missing values as they are (Rusdah & Murfi, 2020), and since our datasets include many missing values whose distribution is not supposed to be random, it is crucial to explicitly handle the missing values of the features. For instance, Fauzan and Murfi (2018) constructed an insurance claim prediction model including XBGoost, Random Forest, and Neural Network models, and they found that the accuracy of XGBoost marks better than any other models after considering missing values.

The XGBoost algorithm is as follows: First, in the initial state \( t = 1 \), there is only a single decision tree. Next, we added a new tree at \( t = 2 \), which shows a smaller loss function. The loss function \( L(t) \) at round \( t \) is expressed as follows:

\[
L(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2} \lambda \|\omega\|^2
\]  

(1)

where \( x_i \) is the \( i \)th municipal and national census datasets and \( y_i \) is the \( i \)th dummy variable whether the house is vacant or not, \( l(\cdot) \) is a differentiable convex loss function representing the error between prediction \( \hat{y}_i \) and actual \( y_i \). We conducted a second-order approximation employing the Taylor series to estimate \( l(\cdot) \). We considered the number of leaf nodes \( T \) and the L2 norm of the weights \( \omega \) to reduce the model complexity and overfitting. \( \gamma \) and \( \lambda \) are the tuning parameters. By minimizing the loss function, we obtain the optimal solution \( w^*_j \) or the \( j \)th leaf node, and derive the minimum value of the loss function. A new tree was added to each round to update the minimum value of the loss function.

We considered the distribution of missing values as not random, but meaningful in itself. For example, if the BR data are missing, it could mean that it is either not inhabited or has no registered residents. Such structures have a higher probability of being vacant than other structures. In XGBoost, instances with missing values were classified in the default direction to allow for a more accurate prediction because missing values in the data can lower the value of the loss function. In addition, this algorithm, which considers sparsity, is faster to compute than the algorithm that does not (Chen & Guestrin, 2016).

Finally, three parameters needed to be tuned: (1) max_depth; (2) subsample; and (3) n_estimators. First, we conducted a grid search by dividing the analyzed data into 70% training data and 30% test data and setting the evaluation value to “f1.”

- max_depth: 4, 5, 6, 7, 8, 9, 10, 11, 12
- subsample: 0.7, 0.75, 0.8, 0.85

Subsequently, we performed machine learning using the optimal parameters obtained. The third parameter, \( n \) _estimators, was tuned by the early stopping process. If the log loss did not decrease for 50 rounds, the learning process is completed. The learning rate and colsample_bytree are fixed respectively at 0.08 and 1, and the ratio of train data to test data is 7:3. We used default values for the other parameters.

### 4.2 Model classification

We categorized our models into 15 combinations. In all, 14 models used a classification framework that included at least one municipal data point for each building unit. Some models have missing values as their geocoded ratios are not 100%. However, XGBoost can perform a significant analysis with missing values. One model used the regression framework, which included only the NC data without building-based information; thus, it could predict the vacant
housing ratio only at the census level. In this model, the supervised data are the vacant housing ratio, which is calculated by aggregating VH data by each basic unit district. Table 3 shows the combination of datasets and the geocoded percentage of samples for the 15 models.

### 4.3 Evaluation metrics

We applied evaluation metrics to determine the practical value of the vacant housing classification model proposed in this study. The confusion matrix is also a standard indicator in the XGBoost classification model (Liu, 2020; Rusdah & Murfi, 2020). First, two indices based on the confusion matrix were used to predict the classification of vacant houses for each unit: recall and precision. In the confusion matrix, vacant houses are represented as positive and non-vacant houses as negative. Here, recall $r_i$ and precision $p_i$ are expressed through the following equations:

$$r_i = \frac{TP}{TP + FN}$$

$$p_i = \frac{TP}{TP + FP}$$

where $TP$, $FN$, and $FP$ indicate true positive, false negative, and false positive, respectively. The former ratio indicates the number of actual vacant houses the model can predict as vacant. The higher its value, the lower the ratio of undetected (overlooked) vacant houses. Therefore, with less oversight, the local government can expect better implementation of their countermeasures against the predicted vacant houses. In contrast, the latter ratio indicates

| Number of data used | Model no. | Data | Geocoded ratio |
|---------------------|-----------|------|----------------|
| 1                   | 1         | ●    | 100%           |
|                     | 2         | ●    | 79.1%          |
|                     | 3         | ●    | 77.3%          |
|                     | 4         | ●    | 76.5%          |
| 2                   | 5         | ●    | 100%           |
|                     | 6         | ●    | 100%           |
|                     | 7         | ●    | 100%           |
|                     | 8         | ●    | 93.7%          |
|                     | 9         | ●    | 88.7%          |
|                     | 10        | ●    | 91.6%          |
| 3                   | 11        | ●    | 100%           |
|                     | 12        | ●    | 100%           |
|                     | 13        | ●    | 100%           |
|                     | 14        | ●    | 97.4%          |
| 4                   | 15        | ●    | 100%           |

Note: “●” in the “Data” represents that the dataset is used in the model. It is true that different geocoded ratios of the datasets change the model tuning parameters, leading to the generation of different models. However, one of the important aspects of this study is how the models derived from the different datasets affect the prediction accuracy considering the different geocoding rates and spatial heterogeneity. Therefore, we perform model tuning using the data selected above.
the number of predicted vacant houses to the number of actual vacant houses. The higher its value, the lower the percentage of occupied houses incorrectly predicted as vacant houses. Thus, local governments can conclude that machine learning prediction is not misguided and implement countermeasures against predicted vacant houses more confidently.

Second, we used the root mean square error (RMSE) as an evaluation index to compare the accuracy of predicting the vacant housing ratio in the “basic unit district.” The following equation expresses the RMSE $e_i$:

$$
e_i = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{y}_k)^2}$$  \hspace{1cm} (4)

where $y_k$ is the actual vacant housing ratio, $\hat{y}_k$ is the predicted vacant housing ratio in the $k$th basic unit district, and the $N$ is the number of basic unit districts. Generally, vacant housing measures cover a specific range of areas. In the target site, several districts specifically required the utilization of vacant houses, which were designated as “priority districts for vacant house measures.” A low RMSE implies a relatively small difference between the vacant housing ratio expected from the machine learning model and the vacant housing ratio investigated from field surveys.

5 | RESULTS

Table 4 shows the results of the metrics and optimal parameters for each of the 15 models. The common point across the models was that the precision is higher than the recall. For all 15 models, the highest values were 0.648 (Model 15), 0.910 (Model 14), and 0.0253 (Model 15) for the recall, precision, and the RMSE, respectively. Thus, the prediction model proposed in this study is more suitable for showing a higher number of predicted vacant houses as actual vacant houses rather than detecting the hidden vacant houses.

Among the models that used a single dataset (Models 1–4), Model 2 showed better results than the other models, whose recall and precision values were 0.168 and 0.582, respectively. This finding is in line with the results of previous studies (Yamashita & Morimoto, 2015) showing that the WU data are useful for detecting vacant houses. In contrast, in Model 1, which used NC data, only the RMSE could be evaluated. The RMSE of Model 1 was 0.079, which shows that the NC data are sufficient to predict a reliable vacant housing ratio per basic unit district. After combining the data, most of the models improved their accuracy.

Models that used two datasets (Models 5–10), for example, Model 8, which used WU and BR data, showed a recall of 0.257, a precision of 0.634, and an RMSE of 0.0706. These values are better than that of models that used WU or BR data individually (Models 2 and 3). Furthermore, Model 10, which used BR and RR data, showed a recall of 0.478, a precision of 0.599, and an RMSE of 0.0551. Interestingly, these values suggest that although the contribution of WU data to the detection of vacant houses is highest when used alone, the contribution of BR and RR data becomes higher by combining them. Thus, a combination of BR and RR data can serve as a substitute for WU data when it is not available.

Among models that used three datasets (Model 11–14), Model 14, which combined the WU, BR, and RR data, showed the highest accuracy for all 15 models, with values of 0.634 for the recall, 0.910 for the precision, and 0.0266 for the RMSE. Model 14 significantly improved the accuracy of the results compared to any other model (Models 8–10) that used two of the three datasets. In contrast, the addition of the NC data did not necessarily improve the accuracy of the two-dataset models (Models 8, 9, and 10) that did not use the NC data. For example, the precision of Model 12 was lower than that of Model 9, and the recall of Model 13 was lower than that of Model 10.

The addition of NC data in Model 15, which used all the municipal datasets, was not effective in improving its accuracy. For instance, compared to Model 14, Model 15 only showed a small increase in recall and RMSE, (0.014 and 0.0013, respectively).
Next, we attempted further improve the interpretability of Model 15, which contained WU, BR, RR, and NC data and showed the best performance among the 15 models. Figure 4a shows Model 15 is classified mainly based on the WU features. However, not every sample contained WU data. Because the XGBoost model can be estimated even in the presence of missing data, it is possible that the model did not effectively classify the samples without the WU data. To address this problem, samples of Model 15 were classified into two groups in terms of whether the WU data of the samples were geocoded (Model 16a) or not (Model 16b). Table 5 presents the results from separating Model 15. Model 16 combined the results of models 16a and 16b.

We then further estimated the two groups using the XGBoost model. By classifying the datasets with and without WU data, Figure 4b confirmed that the WU data includes crucial features such as the average use of water in 2017 and the month of hydrant closure. Meanwhile, Figure 4c shows that features regarding BU data, such as total floor area and building age function, are important.

Table 6 presents the results of the additional estimates. While Models 15 and 16 had similar overall accuracies, Models 16a and 16b had significant differences. The sample without geocoded WU data (Model 16b) showed higher accuracy compared to the samples whose WU data were geocoded (Model 16a). In general, the lower the geocoded ratio and the more missing values, the more difficult it would be to predict vacant housing. However, this result shows the opposite; the low geocoded ratio would contribute to the accuracy of the models.

The geocoded ratio consists of the performance of the geocoding service and the characteristics of the data, but distinguishing the two factors is difficult. The geocoding service we used is one of the services with the highest accuracy in Japan, allowing address matching at a lot identification level, and we tried to eliminate inaccurate matching by preprocessing the address list as much as possible. Therefore, we presume that the ratio of the geocoded samples depends on the characteristics of the data, rather than the performance of the geocoding service. The result indicates...
that the samples with missing values of WU data tend to be vacant houses rather than the ones with specific values. Hence, we need to regard the missing values as the important contributors to detect vacant houses.

To further discuss the difference between Models 16a and 16b, we analyzed the spatial distribution of the predicted values. Figure 5 shows the geocoded ratio of WU data, and it confirmed that a geocoded ratio is high in suburbs and low in the city center and far-suburbs. Hence, while the samples of Model 16a are largely located in suburbs, the ones of Model 16b is in the city center and far-suburbs.

We calculated the mean absolute error (MAE) $e_j$ to understand the spatial distribution of the model accuracies, using the following equation:

$$ e_j = \frac{1}{l} \sum_{i=1}^{l} |m_{ij} - \hat{m}_{ij}| $$

FIGURE 4 The 10 most important features of models: (a) Model 15; (b) Model 16a (for samples with geocoded WU data); and (c) Model 16b (for samples without geocoded WU data). The feature importance implies the number of times the feature is used to split data. “Agg.” implies that the feature is the value aggregated in a basic unit district. Italicized texts show the feature’s data source.

TABLE 5 Division of model 15 by WU data

| Model no. | Number of samples | Number of vacant houses | Vacant housing rate |
|-----------|-------------------|-------------------------|---------------------|
| 15        | 88,363            | 4494                    | 0.0509              |
| 16a       | 69,915            | 1744                    | 0.0249              |
| 16b       | 18,448            | 2750                    | 0.149               |

Note: Model 16a contains samples with geocoded WU data. Model 16b contains samples without geocoded WU data.
where \( m_{ij} \) represents the \( i \)th actual vacant house dummy, \( \hat{m}_{ij} \) is the \( i \)th predicted vacancy dummy for whether a building is vacant (=1) or not (=0), and \( I_j \) is the number of buildings in \( j \)th 500-m square grid. We used 500-m square grids for calculating the MAE instead of basic unit districts, because grid cells evenly divide the target area, which made it convenient to compare the spatial differences without considering the areal weights. In addition, although it is possible to aggregate the prediction per 250 m grid, we chose a 500 m grid for the aggregation scale due to the protection of personal information.

First, Figure 6 shows that grid cells in the urban centers, coastal villages, and mountainous areas marked high MAE values for all the models. These high values can be associated with the number of the samples in each grid. For example, if the number of the samples in the grid were two, only one error would yield an MAE of 0.5. Moreover, other factors might have been more relevant than the missing WU data. For detecting vacant houses with higher accuracy using Model 16, more data such as the economic status of residents would be necessary.

Second, the MAE of Model 16a is high in the city center and far-suburbs, while that of Model 16b is low in the whole area. This finding illustrates differences between the samples with and without geocoded WU data. High MAE

**TABLE 6** Result of the additional estimation

| Model no. | Metrics | Tuned parameter |
|-----------|---------|-----------------|
|           | Recall  | Precision | RMSE | Max_depth | Subsample | n_estimators |
| 16a       | 0.645   | 0.900     | 0.0254 | 4         | 0.85      | 119          |
| 16a       | 0.329   | 0.683     | 0.0288 | 4         | 0.8       | 116          |
| 16b       | 0.825   | 1.00      | 0.0236 | 4         | 0.85      | 65           |
| 16        | 0.633   | 0.914     | 0.0245 | –         | –         | –            |

Note: A precision of 1.00 in Model 16b implies that every sample predicted to be vacant is an actual vacant house. The metrics for model 16 were calculated by summing each value of the confusion matrices for Models 16a and 16b.

**FIGURE 5** Geocoded ratio of WU data.
values of Model 16a in the city center and far-suburbs may be attributed to the disparity in sample size across the city. In other words, spatial heterogeneity of the samples consequently builds a model that fits to a specific area like suburbs, and decreases the predictive accuracy of the other areas. On the other hand, Model 16b (sample without geocoded WU data) showed low values of MAE even in suburbs, despite having relatively small samples in the area. It is conceivable that the model is not overfitted to a particular area and the feature importance, such as building age, gross floor area, maximum age in a household, and so forth, is not spatially weighted significantly across the city.

6  |  CONCLUSION

This study contributes to the few studies on how to determine the spatial distribution of vacant houses quickly and inexpensively, even when there are data accessibility constraints for closed data held by local governments. The 16 models classified in this study were machine-learning-based models using a common framework. Their evaluation and comparison revealed how the prediction accuracy of the models changed depending on the combination of the datasets used. This study has three main contributions.

First, we showed quantitative results for the question of which data were the most relevant to the vacant house prediction model. In the cases where the data are used alone (Models 1–4), the model using WU data (Model 2) has the highest accuracy. However, in cases where two datasets are combined and only one dataset is missing (Models 11–14), the results suggest that RR data and BR data are more important for accuracy than the NC data. Compared to the other models, the model lacking RR data (Model 11) has a larger RMSE, while the model lacking BR data (Model 12) has lower recall and precision. Second, it is possible to acquire a similar accuracy by employing other data, even when specific municipal data are not available. For example, if the WU data are not available, but the BR data are available, Model 6 can be used instead of Model 5. Both models showed almost similar precision; moreover, Model 6 had a lower RMSE. Third, whether WU data are geocoded can significantly affect recall. The previous literature regarded samples with WU Data marked as "out of service" or "inactive" as vacant houses (Yamanshita & Morimoto, 2015). However, this study suggests that rather than the value, the availability of WU data is more important. This finding is unique to XGBoost, which can handle missing values.

This study presents the following challenges for achieving sufficient classification accuracy. Each of these challenges may correlate with each other, resulting in the models’ diminishing utility.

- Overall, the models had a low recall, implying that many vacant houses were left out of the models’ prediction of vacant houses. To detect vacant houses more comprehensively, the models’ recall must be improved using a different machine learning approach on samples without geocoded data.
- For some datasets, the data geocoded rate was not high enough. There is a possibility that the important attributes had missing values and the model did not learn them correctly. To solve this problem, either the geocoding method needs to be improved, or machine learning needs to be performed separately for samples with and without geocoded data.
- Open data were not fully utilized. Although the open NC data used in this study seemed to be useful in predicting the vacant housing rate, it did not significantly improve the models’ accuracy when combined with other datasets (in fact, it sometimes complicated the model and reduced model accuracy). However, in a site where data availability is severely limited, open data can be useful for estimating the vacant housing ratio of each district.

To address the above challenges, future research could examine the applicability of the method in other cities. Although municipalities with comparable sizes and characteristics as the Wakayama City are expected to show similar results, large-sized municipalities in Tokyo’s suburbs or small-sized ones in mountainous areas would have different factors contributing toward the increase in vacant houses. Therefore, a future study uses the models of this study to apply data from other cities to analyze the over-fitting of the models. Such research would enable more local govern-
FIGURE 6  Spatial distribution of MAE per 500-m square grid: (a) Model 15; (b) Model 16a (for samples with geocoded WU data); (c) Model 16b (for samples without geocoded WU data); and (d) Model 16 (for every sample).
ments to understand the distribution of vacant houses with greater accuracy and less effort. Lastly, although access to municipal data is currently limited to researchers in Japan, any municipality holds their data and is potentially able to provide them to other municipalities, which has a potentiality that the method can be deployed in all municipalities.

ACKNOWLEDGMENTS
This research was the result of the joint research with CSIS, the University of Tokyo (No. 880) and used the following data: Zmap TOWN II 2020 Shapefile Wakayama prefecture provided by Zenrin Co., Ltd. We would also like to thank Wakayama Data Utilization Promotion Center and Wakayama city for the provision of the original dataset.

CONFLICT OF INTEREST
The authors declare no conflicts of interest associated with this manuscript.

DATA AVAILABILITY STATEMENT
Research data are not shared.

ORCID
Hiroki Baba https://orcid.org/0000-0003-2005-4578

REFERENCES
Abe, S., Nakagawa, D., Matsunaka, R., & Oba, T. (2014). Study on the factors to transform underused land focusing on the influence of railway stations in central areas of Japanese local cities. Land Use Policy, 41, 344–356. https://doi.org/10.1016/j.landusepol.2014.06.021
Accordino, J., & Johnson, G. T. (2000). Addressing the vacant and abandoned property problem. Journal of Urban Affairs, 22, 301–315. https://doi.org/10.1111/0735-2166.00058
Akiyama, Y., Ueda, A., Ouchi, K., Ito, N., Ono, Y., Takaoka, H., & Hisadomi, K. (2019). Estimating the spatial distribution of vacant houses using public municipal data. In P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, & A. Monsourian (Eds.), Geospatial technologies for local and regional development (pp. 165–183). Springer.
Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science, 355, 483. https://doi.org/10.1126/science.aal4321
Baba, H., Akiyama, Y., Tokudomi, T., & Takahashi, Y. (2020). Learning geographical distribution of vacant houses using closed municipal data: A case study of Wakayama city, Japan. ISPRS Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences, 6(4/W2), 1–8. https://doi.org/10.5194/isprs-annals-VI-4-W2-2020-1-2020
Baba, H., & Asami, Y. (2017). Regional differences in the socio-economic and built-environment factors of vacant house ratio as a key indicator for spatial urban shrinkage. Urban and Regional Planning Review, 4, 251–267. https://doi.org/10.14398/urpr.4.251
Bernt, M. (2019). The emergence of “Stadtumbau ost”. Null, 40, 174–191. https://doi.org/10.1080/02723638.2017.1332926
Bontje, M. (2004). Facing the challenge of shrinking cities in East Germany: The case of Leipzig. GeoJournal, 61, 13–21. https://doi.org/10.1007/s10708-005-0843-2
Bourne, J. (2019). Empty homes: Mapping the extent and value of low-use domestic property in England and Wales. Palgrave Communications, 5, 9. https://doi.org/10.1057/s41599-019-0216-y
Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C., & Liu, R. (2019). XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access, 7, 13149–13158. https://doi.org/10.1109/ACCESS.2019.2893448
Cheshire, P., Hilber, C. A. L., & Koster, H. R. A. (2018). Empty homes, longer commutes: The unintended consequences of more restrictive local planning. Journal of Public Economics, 158, 126–151. https://doi.org/10.1016/j.jpubeco.2017.12.006
Couch, C., & Cocks, M. (2013). Housing vacancy and the shrinking city: Trends and policies in the UK and the city of Liverpool. Null, 28, 499–519. https://doi.org/10.1080/02673037.2013.760029
Du, M., Wang, L., Zou, S., & Shi, C. (2018). Modeling the census tract level housing vacant housing ratio with the Jilin1-03 satellite and other geospatial data. Remote Sensing, 10, 1920. https://doi.org/10.3390/rs10121920
Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M. (2012). Perspectives on big data and big data analytics. Database Systems Journal, 3(4), 3–14. https://EconPapers.repec.org/RePEc:aes:dbjour:v:3:y:2012:i:4:p:3-14
Yamanshita, S., & Morimoto, A. (2015). Study on occurrence pattern of the vacant houses in the local hub city. Journal of the City Planning Institute of Japan, 50, 932–937. https://doi.org/10.11361/journalcpij.50.932
Zabel, J. (2016). A dynamic model of the housing market: The role of vacancies. The Journal of Real Estate Finance and Economics, 53, 368–391.
Zahirovich-Herbert, V., & Gibler, K. M. (2014). The effect of new residential construction on housing prices. Journal of Housing Economics, 26, 1–18. https://doi.org/10.1016/j.jhe.2014.06.003

How to cite this article: Sayuda, K., Hong, E., Akiyama, Y., Baba, H., Tokudomi, T., & Akatani, T. (2022). Accuracy of vacant housing detection models: An empirical evaluation using municipal and national census datasets. Transactions in GIS, 26, 3003–3027. https://doi.org/10.1111/tgis.12992
## DESCRIPTIVE STATISTICS OF FEATURES

| Data | Attribute | Feature Description | Mean | SD | Min | Median | Max |
|------|-----------|---------------------|------|----|-----|--------|-----|
| VH   | Vacant house dummy | | 0.051 | 0.22 | 0 | 0 | 1 |
| LU   | Zoning district | Exclusive residential area dummy | 0.114 | 0.318 | 0 | 0 | 1 |
|      |            | Residential area dummy | 0.115 | 0.319 | 0 | 0 | 1 |
|      |            | Commercial area dummy | 0.027 | 0.163 | 0 | 0 | 1 |
|      |            | Industrial area dummy | 0.022 | 0.148 | 0 | 0 | 1 |
|      |            | Building coverage ratio | 61.034 | 7.122 | 30 | 60 | 80 |
|      |            | Designated floor area ratio | 203.427 | 58.443 | 50 | 200 | 600 |
| RR   |            | Maximum age in a household | 65.665 | 16.097 | 17 | 68 | 109 |
|      |            | Minimum age in a household | 42.559 | 26.438 | 0 | 43 | 108 |
|      |            | Number of residents | 2.932 | 1.769 | 1 | 3 | 19 |
|      |            | Resident register dummy | 0.765 | 0.424 | 0 | 1 | 1 |
| BR   |            | Housing dummy | 0.973 | 0.163 | 0 | 1 | 1 |
|      | Building structure | Wood structure dummy | 0.712 | 0.453 | 0 | 1 | 1 |
|      |            | Steel structure dummy | 0.161 | 0.367 | 0 | 0 | 1 |
|      |            | Reinforced concrete structure dummy | 0.038 | 0.192 | 0 | 0 | 1 |
|      |            | Building age | 28.79 | 14.432 | 0 | 30 | 94 |
|      |            | Gross floor area | 110.983 | 47.669 | 10.07 | 104.49 | 999 |
|      |            | Number of stories | 1.909 | 0.36 | 1 | 2 | 6 |
|      |            | Building registration dummy | 0.773 | 0.419 | 0 | 1 | 1 |
| WU   | Average water use | Average water use in 2018 | 217.995 | 345.886 | 0 | 209 | 55355 |
|      |            | Average water use in 2017 | 214.798 | 364.351 | 0 | 205 | 63373 |
|      |            | Whether the hydrant is open | 0.95 | 0.218 | 0 | 1 | 1 |
|      |            | Number of months since the hydrant was closed | 3.753 | 23.394 | 0 | 0 | 349.5 |
|      |            | Water usage dummy | 0.791 | 0.406 | 0 | 1 | 1 |
| NC   | Type of household | General households | 67.865 | 71.392 | 0 | 46 | 861 |
|      |            | Single person | 10.036 | 10.622 | 0 | 7 | 117 |
|      |            | Students in dormitories | 0.01 | 0.366 | 0 | 0 | 18 |
|      |            | Inmates of medical facilities | 0.152 | 4.143 | 0 | 0 | 241 |
|      |            | Inmates of social institutions | 0.438 | 3.488 | 0 | 0 | 69 |
|      |            | Other types of households | 0.004 | 0.089 | 0 | 0 | 3 |
|      |            | Residents of correctional facilities | 0 | 0 | 0 | 0 | 0 |
|      |            | Type of residence | 60.403 | 65.816 | 0 | 40 | 884 |
|      |            | Row house | 2.567 | 6.995 | 0 | 0 | 116 |
|      |            | Apartment | 14.176 | 28.291 | 0 | 0 | 376 |
|      |            | Other | 0.152 | 1.049 | 0 | 0 | 45 |
| Data Attribute                             | Feature          | Mean  | SD     | Min | Median | Max  |
|-------------------------------------------|------------------|-------|--------|-----|--------|------|
| Number of stories                         | 1–2              | 6.422 | 14.203 | 0   | 0      | 120  |
|                                           | 3–5              | 4.89  | 15.388 | 0   | 0      | 256  |
|                                           | 6–10             | 2     | 12.522 | 0   | 0      | 306  |
|                                           | 11–14            | 0.759 | 11.061 | 0   | 0      | 366  |
|                                           | 15–              | 0.105 | 3.486  | 0   | 0      | 158  |
|                                           | 1–2              | 9.61  | 17.792 | 0   | 0      | 181  |
|                                           | 3–5              | 3.405 | 10.177 | 0   | 0      | 129  |
|                                           | 6–10             | 0.988 | 6.997  | 0   | 0      | 148  |
|                                           | 11–14            | 0.166 | 2.463  | 0   | 0      | 90   |
|                                           | >15              | 0.007 | 0.219  | 0   | 0      | 9    |
| House floor number                        | 1–2              | 60.357| 65.611 | 0   | 40     | 865  |
|                                           | 3–5              | 1.257 | 10.709 | 0   | 0      | 256  |
|                                           | 6–10             | 14.143| 22.153 | 0   | 5      | 242  |
|                                           | 11–14            | 1.037 | 4.842  | 0   | 0      | 224  |
|                                           | >15              | 0.503 | 1.346  | 0   | 0      | 108  |
| Type of housing and ownership             | Owned house      | 6.705 | 8.127  | 0   | 4      | 111  |
|                                           | Public rented house | 0.364 | 1.255  | 0   | 0      | 12   |
|                                           | Public corporation's rented house | 1.203 | 2.18   | 0   | 0      | 18   |
|                                           | Private rented house | 1.574 | 4.933  | 0   | 0      | 44   |
|                                           | Residence for employees | 0.242 | 0.91   | 0   | 0      | 8    |
|                                           | Room rent        | 0.177 | 1.215  | 0   | 0      | 16   |
|                                           | Company dormitories | 0.067 | 0.798  | 0   | 0      | 18   |
| Number of residents in a household        | 1                | 10.038| 10.622 | 0   | 7      | 117  |
|                                           | 2                | 20.46 | 19.05  | 0   | 14     | 168  |
|                                           | 3                | 18.44 | 20.503 | 0   | 12     | 246  |
|                                           | 4                | 18.739| 25.596 | 0   | 12     | 348  |
|                                           | 5                | 7.104 | 10.28  | 0   | 5      | 155  |
|                                           | 6                | 2.241 | 4.431  | 0   | 0      | 42   |
|                                           | 7                | 0.672 | 2.367  | 0   | 0      | 21   |
|                                           | 8                | 0.177 | 1.215  | 0   | 0      | 16   |
|                                           | 9                | 0.067 | 0.798  | 0   | 0      | 18   |
| Type of family members                    | Married couple only | 14.906| 14.642 | 0   | 10     | 134  |
|                                           | Married couple with child/children | 34.234| 44.449 | 0   | 20     | 560  |
|                                           | Male parent and child/children | 1.008 | 1.795  | 0   | 0      | 17   |
|                                           | Female parent and child/children | 6.705 | 8.127  | 0   | 4      | 111  |
|                                           | Couple and parents | 0.364 | 1.255  | 0   | 0      | 12   |
|                                           | Couple and single parent | 1.203 | 2.18   | 0   | 0      | 18   |
|                                           | Couple, children, and parents | 1.574 | 4.933  | 0   | 0      | 44   |
|                                           | Couple and other relatives | 0.242 | 0.91   | 0   | 0      | 8    |
|                                           | Couple, children, and other relatives | 1.16  | 2.6    | 0   | 0      | 21   |
|                                           | Couple, parents, and other relatives | 0.157 | 0.942  | 0   | 0      | 15   |
| Data Attribute | Feature | Number of children |
|----------------|---------|--------------------|
|                |         | 0                  |
|                |         | 1                  |
|                |         | 2                  |
|                |         | 3                  |
|                |         | >4                 |
| Type of households with residents aged 65 years and above | Couple, children, parents, and other relatives | 0.601 | 2.067 | 0 | 0 | 20 |
| | Siblings only | 0.343 | 0.859 | 0 | 0 | 8 |
| | Other | 1.37 | 2.509 | 0 | 0 | 23 |
| | Including non-relatives | 0.622 | 1.536 | 0 | 0 | 12 |
| | Single household | 10.036 | 10.622 | 0 | 7 | 117 |
| | Unknown | 0.177 | 0.879 | 0 | 0 | 12 |
| | Only 65 years and above; one-person household | 4.722 | 4.383 | 0 | 4 | 67 |
| | Only 65 years and above; two-person household; couple only | 7.895 | 7.742 | 0 | 6 | 60 |
| | Only 65 years and above; two-person household; others | 0.344 | 0.852 | 0 | 0 | 6 |
| | Only 65 years and above; household of three or more persons; couple and their parents | 0.249 | 0.907 | 0 | 0 | 6 |
| | Households with three or more members; others | 0.079 | 0.487 | 0 | 0 | 4 |
| | Including those under 65; two-person household; couple only | 1.62 | 2.467 | 0 | 0 | 26 |
| | Including under 65; household of two persons; other | 2.76 | 3.106 | 0 | 2 | 24 |
| | Including under 65; household of three or more persons; Son and his wife | 4.77 | 7.007 | 0 | 3 | 67 |
| | Including under 65; household of three or more persons; daughter and her husband | 1.463 | 3.041 | 0 | 0 | 27 |
| | Including under 65; household of three or more persons; single child | 0.984 | 2.139 | 0 | 0 | 17 |
| | Including under 65; household with three or more persons; son only | 3.288 | 4.259 | 0 | 3 | 31 |
| | Including under 65; household of three or more persons; daughter only | 2.476 | 3.54 | 0 | 0 | 28 |
| | Including under 65; household of three or more persons; other | 2.225 | 3.591 | 0 | 0 | 28 |

Number of children
| Data                        | Attribute                        | Feature            | Mean  | SD    | Min | Median | Max |
|-----------------------------|----------------------------------|--------------------|-------|-------|-----|--------|-----|
| Maximum age of the children |                                   | 0                  | 2.126 | 4.787 | 0   | 0      | 69  |
|                             |                                  | 1–2                | 3.949 | 7.734 | 0   | 0      | 140 |
|                             |                                  | 3–5                | 4.44  | 8.112 | 0   | 0      | 113 |
|                             |                                  | 6–8                | 4.026 | 7.396 | 0   | 0      | 105 |
|                             |                                  | 9–11               | 3.735 | 6.81  | 0   | 0      | 111 |
|                             |                                  | 12–14              | 4.018 | 7.153 | 0   | 0      | 111 |
|                             |                                  | 15–17              | 3.836 | 6.733 | 0   | 0      | 136 |
|                             |                                  | >18                | 22.555| 23.303| 16  | 0      | 328 |
| Minimum age of the children | 0                                | 0.816              | 2.286 | 0     | 0   | 36     |
|                             | 1–2                              | 1.843              | 4.043 | 0     | 0   | 65     |
|                             | 3–5                              | 3.161              | 6.565 | 0     | 0   | 93     |
|                             | 6–8                              | 3.577              | 7.241 | 0     | 0   | 129    |
|                             | 9–11                             | 3.707              | 6.759 | 0     | 0   | 101    |
|                             | 12–14                            | 4.214              | 7.619 | 0     | 0   | 129    |
|                             | 15–17                            | 4.772              | 8.356 | 2     | 0   | 136    |
|                             | >18                              | 26.595             | 28.518| 0     | 18  | 478    |
| Sex of the residents        | Male                             | 37.218             | 37.596| 0     | 26  | 448    |
|                             | Female                           | 41.287             | 40.451| 0     | 30  | 464    |
| Marital status              | Unmarried                        | 26.632             | 29.384| 0     | 18  | 359    |
|                             | With spouse                      | 39.93              | 41.159| 0     | 27  | 492    |
|                             | Bereaved                         | 6.388              | 5.698 | 0     | 5   | 53     |
|                             | Divorced                         | 4.245              | 4.774 | 0     | 3   | 45     |
|                             | Unknown                          | 1.31               | 2.526 | 0     | 0   | 41     |
| Nationality                 | Japan                            | 77.805             | 77.097| 0     | 55  | 901    |
|                             | Foreign country                  | 0.474              | 1.249 | 0     | 0   | 18     |
|                             | Unknown                          | 0.226              | 0.868 | 0     | 0   | 16     |
| Period of residence         | From the time of birth           | 9.468              | 10.596| 0     | 6   | 134    |
|                             | <1                               | 3.568              | 6.616 | 0     | 1   | 130    |
|                             | 1–5                              | 10.997             | 17.052| 0     | 5   | 281    |
|                             | 5–10                             | 9.727              | 15.625| 0     | 5   | 254    |
|                             | 10–20                            | 13.76              | 21.44 | 0     | 7   | 373    |
|                             | >20                              | 26.937             | 25.803| 0     | 19  | 307    |
| Place of residence before 5 years | Current address                  | 60.809             | 59.961| 0     | 43  | 812    |
|                             | Within the same district         | 9.254              | 15.126| 0     | 4   | 233    |
|                             | Another district in the same prefecture | 1.605         | 3.399 | 0     | 0   | 44     |
|                             | Another prefecture               | 2.669              | 5.223 | 0     | 1   | 82     |
|                             | Abroad                           | 0.11               | 0.615 | 0     | 0   | 12     |

(Continues)
| Data         | Attribute                          | Feature                        | Mean  | SD      | Min | Median | Max |
|-------------|------------------------------------|--------------------------------|-------|---------|-----|--------|-----|
| Work status | Employed                           | 35.624                         | 37.963| 0       | 24  | 508    |     |
|             | Unemployed                         | 1.745                          | 2.189 | 0       | 1   | 28     |     |
|             | Housework                          | 11.069                         | 10.551| 0       | 8   | 101    |     |
|             | Educated                           | 9.174                          | 13.076| 0       | 5   | 191    |     |
|             | Other                              | 17.31                          | 16.285| 0       | 13  | 267    |     |
| Commute status | Neither employed nor educated     | 30.124                         | 27.08 | 0       | 22  | 267    |     |
|             | Work at home                       | 3.34                           | 3.916 | 0       | 2   | 53     |     |
|             | Within the same district           | 34.103                         | 39.406| 0       | 22  | 513    |     |
|             | To another district in the same prefecture | 2.72  | 4.897 | 0       | 1   | 109    |     |
|             | To another prefecture              | 2.954                          | 4.107 | 0       | 2   | 62     |     |
|             | Unknown                            | 0.136                          | 0.435 | 0       | 0   | 5      |     |
| Employment status | Regular staff and employees | 18.282                         | 21.521| 0       | 12  | 280    |     |
|             | Temporary employees                | 0.558                          | 0.949 | 0       | 0   | 8      |     |
|             | Part-time workers                  | 9.341                          | 11.118| 0       | 6   | 154    |     |
|             | Officers                           | 1.731                          | 2.186 | 0       | 1   | 23     |     |
|             | Employer with employees            | 0.946                          | 1.284 | 0       | 1   | 11     |     |
|             | Self-employed                      | 2.48                           | 2.797 | 0       | 2   | 25     |     |
|             | Family employees                   | 1.2                            | 1.853 | 0       | 1   | 32     |     |
|             | Domestic workers                   | 0.062                          | 0.278 | 0       | 0   | 3      |     |
| Type of industry | Agriculture                      | 0.628                          | 2.271 | 0       | 0   | 44     |     |
|             | Forestry                           | 0.011                          | 0.139 | 0       | 0   | 4      |     |
|             | Fishery                            | 0.058                          | 0.526 | 0       | 0   | 15     |     |
|             | Mining, quarrying, and gravel extraction | 0.005 | 0.068 | 0       | 0   | 1      |     |
|             | Construction                       | 2.725                          | 3.603 | 0       | 2   | 39     |     |
|             | Manufacturing                      | 5.428                          | 7.266 | 0       | 3   | 115    |     |
|             | Electricity, gas, heat, and water supply | 0.245 | 0.776 | 0       | 0   | 18     |     |
|             | Information and communication      | 0.514                          | 0.951 | 0       | 0   | 13     |     |
|             | Transportation and postal services  | 1.891                          | 2.684 | 0       | 1   | 32     |     |
|             | Wholesale and retail trade         | 5.778                          | 6.473 | 0       | 4   | 84     |     |
|             | Finance and insurance              | 1.016                          | 1.649 | 0       | 0   | 22     |     |
|             | Real estate and goods leasing      | 0.635                          | 1.108 | 0       | 0   | 10     |     |
|             | Academic research, technical services | 0.954 | 1.419 | 0       | 0   | 13     |     |
|             | Accommodation and food services    | 1.923                          | 2.67  | 0       | 1   | 27     |     |
|             | Lifestyle services and entertainment | 1.225 | 1.75  | 0       | 1   | 21     |     |
|             | Education and learning support     | 1.778                          | 2.448 | 0       | 1   | 29     |     |
|             | Medical care and welfare           | 5.248                          | 6.605 | 0       | 3   | 84     |     |
|             | Composite service business         | 0.238                          | 0.585 | 0       | 0   | 7      |     |
| Data Attribute | Feature                                      | Mean  | SD    | Min | Median | Max |
|----------------|----------------------------------------------|-------|-------|-----|--------|-----|
| Averaged indices about house and household | Services (not classified elsewhere)          | 2.102 | 2.608 | 0   | 1      | 31  |
|                | Other public services                        | 1.536 | 2.549 | 0   | 1      | 74  |
|                | Industries not elsewhere classified          | 1.685 | 2.338 | 0   | 1      | 21  |
|                | Number of stories                            | 1.433 | 1.963 | 0   | 0      | 15.1|
|                | Floor number of the house                    | 0.982 | 1.246 | 0   | 0      | 9.4 |
|                | Number of residents in a household           | 3.057 | 2.649 | 0   | 2.9    | 114.1|
|                | Number of males in a household               | 1.435 | 1.349 | 0   | 1.4    | 67.2|
|                | Number of females in a household             | 1.623 | 1.488 | 0   | 1.5    | 46.9|
|                | Age                                          | 49.741| 10.119| 0   | 50.1   | 93  |