Original Article

Support system for road price formulation using machine learning

Yunhao Tu (Graduate School of Informatics, Nagoya University, tu.yunhao.p5@s.mail.nagoya-u.ac.jp, Japan)
Mayu Urata (Graduate School of Informatics, Nagoya University, mayu@i.nagoya-u.ac.jp, Japan)
Mamoru Endo (Graduate School of Informatics, Nagoya University, endo@i.nagoya-u.ac.jp, Japan)
Takami Yasuda (Graduate School of Informatics, Nagoya University, yasuda@i.nagoya-u.ac.jp, Japan)
Hirokazu Shimazaki (System Development Department, Japan Appraisal System INC, shimazaki@jasinc.co.jp, Japan)
Tomoyuki Kimura (System Development Department, Japan Appraisal System INC, tomo-kimura@jasinc.co.jp, Japan)

Abstract

Fixed asset tax is assessed based on land use and is an essential financial resource for Japanese municipalities. Therefore, the evaluation of fixed asset tax is important for local governments. A real estate appraiser uses the road price index to determine the fixed asset tax on land; therefore, the appropriate road price should be specified. The local government typically decides the road price based on expert experience and knowledge. Hence, there is need for an objective and transparent basis for road price decisions, and municipal officials currently devote considerable time and effort to surveys for appropriate evaluation. In this study, large-scale data, such as public and private sector data, with machine learning, are utilized to construct a road price estimation system under an industry–government–academia collaboration. An ensemble model utilizing a machine-learning algorithm was applied, i.e., the gradient boosting decision tree, to learn previous road price data and the various factors that affect road prices. In addition, via visualization, the importance of each element that affects road prices was quantified, thereby enabling the determination of the effect of each element or feature objectively. Our proposed system can be used to develop a road price estimation model and predict road prices for new roads. Its usage is aimed at promoting data utilization in local governments to reduce manual labor as well as to improve efficiency and transparency in formulating road prices. Furthermore, it creates new possibilities for the activation and utilization of public and private sector data, such as fixed asset data.

Key words
fixed asset tax, road price formulation, machine learning, gradient boosting decision tree, data utilization

1. Introduction

Information and communication technology (ICT), such as artificial intelligence (AI), has been actively practiced in various fields in recent years. Moreover, digitalization has progressed throughout society such that, owing to the development of the Internet of Things and digital transformation, all online activities can now be recorded as digital data. Furthermore, improvements in computing hardware performance and the development of AI technologies have enabled massive amounts of various types of data to be processed at high speed. As a result, data-centric business models are multiplying, and value creation through data utilization is being promoted in various fields in society; for example, in administration, the importance of data utilization is increasing. In Japan, a declining birthrate and an aging population are indicative that the data possessed by local governments should be utilized to craft evidence-based policies based on data analysis to maintain and improve the quality of resident services with limited financial resources. Herein, methods of effectively applying and thereby promoting data utilization in local governments are discussed.

In Japanese municipalities, fixed asset tax is a basic tax (with a slight bias) that is vital because it exists in every municipality. Figure 1 shows a breakdown of municipal taxes in fiscal year (FY) 2019. The fixed asset tax is 9,286 billion yen, which accounts for 40.6 % of the total tax revenue (Ministry of Internal Affairs and Communications, Japan, 2021a). The fixed asset tax on land is based on the parcel unit, which refers to one land in the land register. The landowner applies for the registered land to record the land in the registry by himself. Meanwhile, municipalities must investigate the current use of a land unit regardless of its registration status, and determine its tax category. This tax payment method is the taxation method. Therefore, reliability and transparency are required when taxpayers use the self-assessment method to calculate and fill in tax amounts. The evaluation of fixed asset tax is projected to become more challenging (Ministry of Internal Affairs and Communications, Japan, 2021b; JRI Review, 2021). By contrast, the decreasing number of local public servants implies that fewer employees will perform administrative tasks in the future. In Handa City, which is the city investigated in this study, eight staff members must examine approximately 110,000 pieces of land in the city (Chita Statistical Research Council, Japan, 2020). Comprehensive investigation of all the lands is difficult. Therefore, we propose a method to solve this issue via industry–government–academic cooperation. Previous land data were utilized, such as road and road prices, to design a system for supporting fixed asset tax evaluation using machine learning.

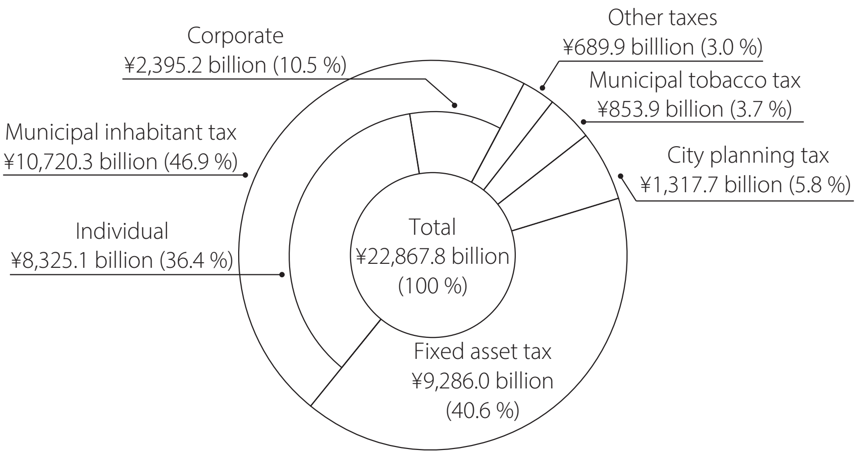


Figure 1: Municipal taxes in FY2019

Source: Created by the authors based on the reference (Ministry of Internal Affairs and Communications, Japan, 2021a).

2. Road price formulation

Local governments use indicators such as road prices, to determine fixed asset tax. A road price indicates the evaluated price per square meter of residential land facing the street (Figure 2) and is reviewed every three years via the following steps:

- Real estate appraisers divide the target area for each purpose. Each divided district is known as a use district, where the "use" includes commercial, residential, and industrial (Figure 3).
- Real estate appraisers divide each use district into smaller groups of areas with different situations. Each divided smaller group is known as a similar situation area (Figure 4).
- Real estate appraisers determine a standard residential land where the price and street conditions are the standards in a similar area. Additionally, an expert sets the street in front of the standard residential land as the "main street" (Figure 5).
- Real estate appraisers appraise the price of standard residential land, where 70 % of the appraised price is the standard price of the main streets (Figure 6).
- Roads other than "main streets" are known as "other streets."

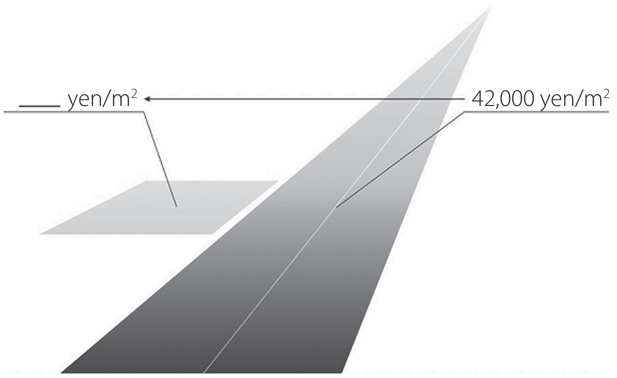


Figure 2: Road price formulation for fixed asset tax

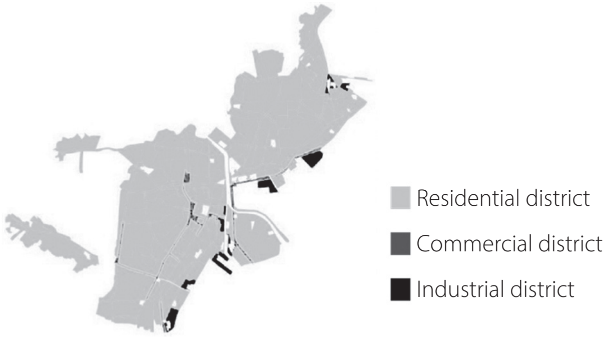


Figure 3: Use districts

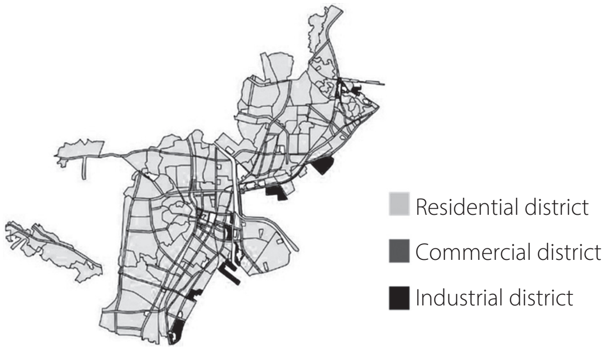


Figure 4: Similar situation areas

Real estate appraisers calculate such a road price by comparing the road with the "main streets" in similar areas to which the road belongs. To calculate the road prices of other streets, experts refer to a comparison table that summarizes the rate of change in road prices due to changes in land price formation factors (Figure 7).

Therefore, real estate appraisers only appraise the road price of the "main street" when calculating the road price. The road prices of other streets are manually calculated based on the comparison table with which experts verify and adjust

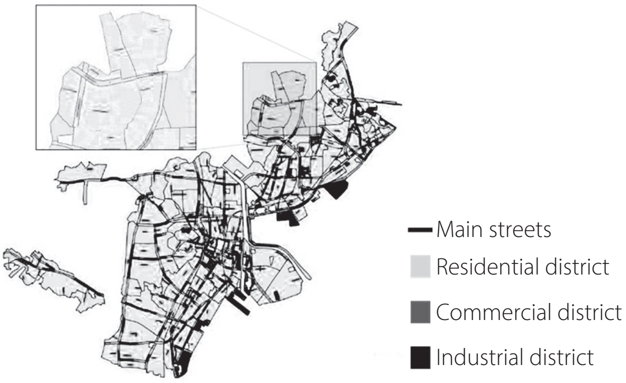


Figure 5: Standard residential land and major streets

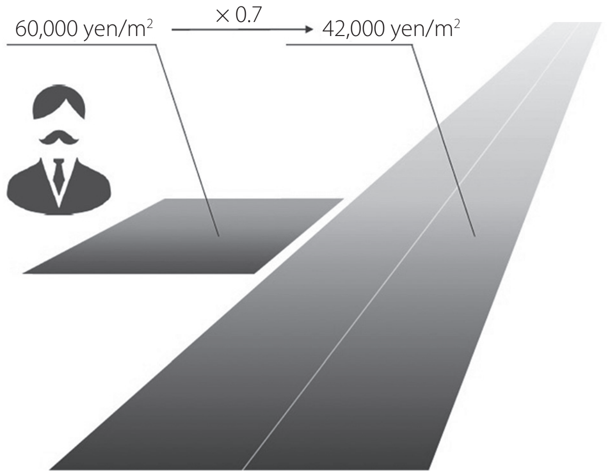
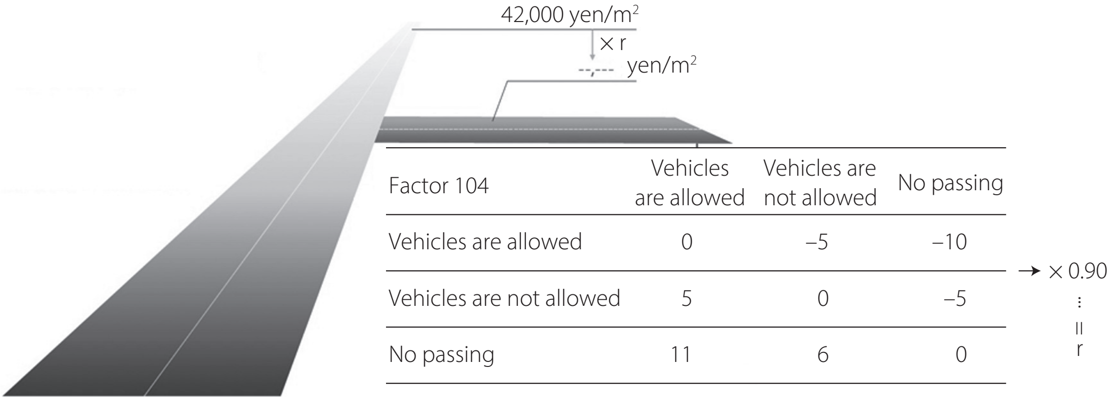


Figure 6: Individual evaluation of standard residential land and land price of main streets

the calculated road prices. Evaluating all routes by individually assessing all roads is difficult, as the number of roads ranges from thousands to tens of thousands. Because the verification and investigation of road prices are arduous, current evaluations are performed based on the experience and knowledge of experts. Therefore, here, a system was constructed to support road price formulation using machine learning based on previous road prices and road data via industry-government-academic cooperation. We decided on this cooperation to compensate for inadequate pool of knowledgeable personnel in the local governments. This system was proposed to reduce the workload of government staff and improve work efficiency.

This study aims to evaluate the effect and significance of using machine learning based on data utilization for local government tasks and improve the understanding of its use for AI. We jointly investigated a land use judgment system using deep learning since 2017 (Ukai et al., 2018; Kato et al., 2019; Tu et al., 2021), from which considerable knowledge and experience were obtained for performing the current road price estimation. The contributions of this study can be summarized as follows, we:

- analyze and summarize the progress and problems of data utilization in Japan and propose a method to solve the problems pertaining to data utilization by local governments using a practical application example.
- investigate the effect of using features for predicting road prices and analyze the effect of learning/not learning external features. The experimental results demonstrate that the inference performance is better in terms of estimation accuracy when no exterior features are learned. Conversely, the inference performance deteriorates after the external features are learned. Nonetheless, the possibility of using publicly available data was investigated, and the results will facilitate future investigations.
- propose a machine learning-based road price formulation system, effectively using public and private sector data from local governments to develop this system over the past three years. Road prices are typically evaluated based on the experience and land knowledge of experts. However, road prices should be evaluated more objectively based on previous road data; therefore, we developed a system that can support local government tasks.



Factor 104	Vehicles are allowed	Vehicles are not allowed	No passing
Vehicles are allowed	0	−5	−10
Vehicles are not allowed	5	0	−5
No passing	11	6	0

Figure 7: Road price evaluation of other streets by comparison table

Further, the proposed system:

- can effectively indicate the relationship between estimated road prices and various land price formation factors via visualization. Furthermore, it can serve as a reference for staff members in the tax department for determining road prices. Visualizing the results allows non-experts to understand the effects of many factors on road prices.
- will allow road prices to be evaluated practically and accurately, and will enable local government staff to perform more efficiently. For example, the system can be used to rapidly estimate the price of new roads, thus reducing the time required to complete tasks.

To facilitate the tasks of local governments, a support system for road price formulation was constructed, based on machine learning using public and private sector data via industry–government–academia collaboration, to be used in fixed asset taxation. The system aims to reduce the complexity required in formulating road prices and improve work efficiency and transparency.

3. Related studies

Land-use judgment and road price estimation are crucial when evaluating fixed asset tax because they are the primary bases for calculating fixed asset tax. This study primarily focuses on road price estimation and introduce studies related to the two aspects in this section, which forms the primary content of our study.

In related studies, machine learning has been used extensively to classify land use and cover (Kussul et al., 2017; Yang et al., 2018). Kussul et al. classified land cover and crop types as well as pixel-level segmentation based on remote sensing data (Kussul et al., 2017). Their model successfully classified major crops with high accuracy. Moreover, the authors visualized the results of image segmentation on a map, which clearly presented the distribution of crops. However, when evaluating fixed asset tax, which is determined primarily on the basis of land use, one must consider the use of agricultural and other lands, such as residential and miscellaneous lands. Therefore, more factors should be considered, such as the diverse use of lands and road prices, to evaluate the fixed asset tax of land comprehensively. Yang et al. investigated a deep learning-based model for classifying land use and land cover in residential areas and urban green spaces (Yang et al., 2018). However, their proposed model only achieved high accuracy in general classifications, i.e., not in specific classifications. The model is not utilized in practical applications because it has not been evaluated in actual scenarios. In contrast, our study involved performing laboratory experiments for specific tasks, such as determining fixed asset tax, performing field investigations, and utilizing the model to construct a system that can be used by local government staff to perform practical tasks.

We have been conducting laboratory studies to evaluate the proposed system based on a real scenario of fixed asset taxation. Beginning in 2017, an initial investigation was conducted into a land judgment system using deep learning for actual land evaluation (Ukai et al., 2018). The system was improved by strengthening its classification accuracy and expanding its classification tasks; furthermore, it was used for field surveys (Tu et al., 2021). Additionally, land evaluation was performed via collaboration with local governments, from which valuable knowledge and experience was obtained. The intention is to apply the experience and expertise gained to investigating road price estimation, the results of which can serve as important reference.

In terms of road price prediction, Li et al. conducted a study to support the verification and adjustment of road prices (Li et al., 2011). To evaluate the results of standard residential land, researchers have predicted the road prices of other streets using a method known as universal kriging, which considers a linear regression model, and structures the spatial correlation of model errors by expressing the covariance of errors between two different points as a distance function. The method enables predictions that consider the spatial factors in road prices. However, further improvements can be realized. First, an average prediction error of 6.9 % between the predicted value and actual road price was recorded in a previous study, indicating that the accuracy can be improved by, for example, improving the estimation method. Second, various land formation factors that affect the setting of road prices and the interaction between each element used to determine road prices should be elucidated. A visualization method can effectively demonstrate the interaction between predicted road prices and various influencing factors is indicated.

Aoki et al. proposed a method using natural classification (Aoki et al., 2016) and attempted to use a GIS to create a road price classification and validation map, which facilitated the visualization of the road prices of municipalities. They applied the proposed method to distribute land prices in large cities, such as Kyoto. Although their findings suggest constructing a road price distribution map, the calculated road prices were approximate, and not an accurate reflection of price changes. Conversely, our proposed method can be used to calculate specific route prices, which can accurately reflect the difference in route prices. By visualizing the results on a map, the situations of local roads can be observed.

Takeda et al. proposed a method for visualizing the distribution and flow of road prices by analyzing land price trends using a natural classification method to improve the efficiency of road price verification (Takeda et al., 2018). After confirming the distribution and trends of the price range, the road price can be efficiently verified by creating a road price flow chart that reflects the specific road price. In contrast, we construct a model that predicts the price of new roads by learning previous road data and the route price. Thus, the efficiency and accuracy of road price formulation can be improved by verifying the road price calculated using the comparison table.

Kawano et al. proposed using machine learning to improve the prediction accuracy of road prices as an improvement on the two previous studies mentioned above (Kawano et al., 2020). Furthermore, they improved the current road price using predicted results. Nonetheless, the prediction accuracy can be further improved. Notably, visualization enables the objective prediction of road prices and allows non-experts to understand the standard and process of road price formulation. The proposed model in this study was used in an actual task to predict the price of a new road.

In addition to the abovementioned studies, researchers have investigated the creation of large-scale datasets for object detection using aerial photographs or satellite images to evaluate fixed asset tax (Yu et al., 2018; Ding et al., 2019; Ding et al., 2021; Xia et al., 2018). Yu et al. used deep learning to analyze satellite images of 48 states in the United States, excluding Alaska and Hawaii, and successfully detected 1.47 million solar panels nationwide (Yu et al., 2018). Therefore, users can view the location and size of solar panels in the United States on a map, thus allowing them to confirm the distribution of solar panels and registration targets. According to their study, the detection model learned approximately 370,000 satellite images. Furthermore, the authors used more than 1 billion satellite images to investigate the distribution of solar panels in 48 states in the United States. Besides this, Ding et al. created and published a large dataset to detect objects in aerial photographs and satellite images (Ding et al., 2019; Ding et al., 2021; Xia et al., 2018). The most recent dataset version of Object deTection in Aerial images (DOTA), i.e., DOTA-v2.0, contains 18 common categories, 11,268 images, and 1,793,658 instances. In their study, the authors obtained images from Google Earth, the GF-2 satellite, and aerial images. However, obtaining such a large-scale dataset is challenging for local governments. Hence, this study uses existing data possessed by local governments and open data published by the government to construct the dataset required for developing the proposed system. This allows the use of existing data and relieves the necessity to obtain new data. Moreover, the study results can be more easily disseminated to other local governments, which promotes its application in local governments across the country and allows the further utilization of fixed asset information.

4. Data utilization

Our study received public and private data, such as previous road data provided by the management staff of Handa City, Aichi Prefecture. Furthermore, open data released by the government were utilized. Details regarding the data are provided below.

4.1 Public and private sector data

The Government of Japan enacted the “Basic Act on the Advancement of Public and Private Sector Data Utilization” to utilize public and private data (Prime Minister of Japan and His Cabinet, Japan, 2016). This law describes public and private sector data and the basic idea. The data referenced here is digital data managed, used, and provided by the government and private businesses to execute tasks. Under this law, prefectures are obliged to formulate a promotion plan for data utilization in public and private sectors. Municipalities, including special administrative regions, are equally obliged to formulate a plan. The objectives of the law stated above include to promote the realization of a vibrant Japanese society and effective administration. Furthermore, it promotes the smooth distribution of information while protecting the rights and interests of individuals and corporations. One of the objectives of this study is to utilize and analyze large-scale data owned by local governments based on this law and provide the results to facilitate administrative tasks.

4.2 Open data

Considering the recycling and convenience of data, most researchers focus on utilizing open data because these can be used secondarily regardless of whether the usage is for commercial purposes, can be used free of charge by anyone, and are machine-readable (IT Strategic Headquarters, Japan, 2012). The government has launched a series of measures to promote use of open data to solve practical problems. “Open Data 2.0” requires promoting open data for problem solving (IT Strategic Headquarters, Japan, 2016). The “Open Data Basic Guidelines” stipulates that the planning, maintenance, and operation of information systems and business processes are to be based on the concept of “open data by design.” The guidelines require the convenient use of data possessed by national and local governments as well as businesses. In principle, data possessed by various departments and agencies are presented as public data (IT Strategic Headquarters, Japan, 2017). Furthermore, the “Declaration to Be the World’s Most Advanced Digital Nation Basic Plan for the Advancement of Public and Private Sector Data Utilization” requires the utilization of ICT and data to transform the social structure (The Government of Japan, 2020).

Therefore, in this study, open data were utilized for local government administrative tasks and analyzing the utilization results. Moreover, providing practical examples will facilitate the utilization of open data by local governments.

4.3 Data utilization and dataset creation

Land data, such as road data and road prices, were used to create a dataset. Management staff from Handa City, Aichi Prefecture, provided three years data, i.e., 2015, 2018, and 2021, which contained 4842 sets of road data, 4964 sets of road data, and 364 sets of primary street road data, respectively. Additionally, 10170 sets of road data were used as training data, while 4604 sets of road data of other streets in 2021 were used as test data. In addition to the factors listed in the comparison table, the road location coordinates, the population density of each town, and the transaction density were considered. The population of each town is based on open data published by the Handa City (Handa City, Japan, 2022). The population density based on the population of each town and the town area was calculated. The number of land transactions per town is based on open data from the Land Information System published by Information System published by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT, Japan, 2022). The model learns previous data and then calculates and verifies the road price based on the test data. Table 1 summarizes the features used for model training and testing.

The effective data utilization method used by local governments was utilized. Specifically, we aim to realize effective data utilization in local governments by utilizing fixed asset information and open data. Furthermore, a support system was developed for performing fixed asset tasks in local governments, reducing workloads, and improving evaluation accuracy and work efficiency. From a theoretical perspective, a method of utilizing data was proposed, such as ICT and fixed asset information, for land evaluation by local governments and aim to promote the further utilization and development of public and private sector data in local governments nationwide.

5. Road price formulation work support system

A system was constructed to support road price formulation via the steps below (Figure 8). Next, the system overview and estimation model are described in detail.

5.1 Gradient boosting decision tree

The gradient boosting decision tree (GBDT) comprises a set of decision trees. It learns and creates an estimation model iteratively to improve the calculated error between the target and predicted values. The predetermined number of decision trees can perform repeated learning. It is a good technique for estimating correct values from multiple factors. The GBDT is often recommended for classification and regression problems involving tabular data. Moreover, GBDT is widely used in online prediction, which plays an important role in many practical industrial applications, such as click prediction in advertising searches, content ranking in web searches, content optimization in recommender systems, and travel time estimation in traffic planning (Ke et al., 2019). In recent years, deep learning has been used for the regression and classification of table data. However, differences in model performance have been reported depending on the dataset used (Shwartz-Ziv and Armon, 2022). In consideration of model stability, the GBDT without deep learning was used.

To estimate road prices, four GBDT models were used, namely XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), and NGBoost (Duan et al., 2020). XGBoost, LightGBM, and CatBoost are classical GBDT models often used in estimating table data. NGBoost can predict probabilistic output values by improving the classical GBDT. Therefore, the abovementioned four models were used to predict land prices in this study. The training data were randomly segregated into eight portions to train the model and the parameters were adjusted using k-fold cross-validation. For k-fold cross-validation, the training data were partitioned into k segments. One segment was used as performance evaluation data, and the remaining segments were used as training data. With this method, learning was repeated k times, and the model performance was evaluated based on each result. When training the model, we first used each model to learn the training data separately and then used the method described above to adjust the initial parameters of the model. Subsequently, we compared and identified the best results by adjusting the random states. To further improve the accuracy of the estimated road prices, we performed the following measures. First, we trained multiple models with different structures using different random states. Next, we calculated the simple average of these models and used them as the output of the model. For example, we trained multiple XGBoost models using different random states. Subsequently, we calculated the simple average of multiple XGBoost models and used them as the output of the XGBoost models. We trained multiple LightGBM, CatBoost, and NGBoost models and obtained simple averages using the same method. Finally, we calculated the simple average of the four models (XGBoost, LightGBM, CatBoost, and NGBoost). The result obtained is the final estimated road price.

5.2 Overview of proposed system

The proposed system comprises the following four steps:

- In step 1, we use large-scale data owned by Handa city to create a dataset for constructing a prediction model of road prices. Subsequently, we segregate the dataset into training and test datasets to construct a road price estimation model. Additionally, we utilize open data published by the government, showing that external data can be used.
- In step 2, we construct a road price estimation model

Table 1: List of features for road price estimation

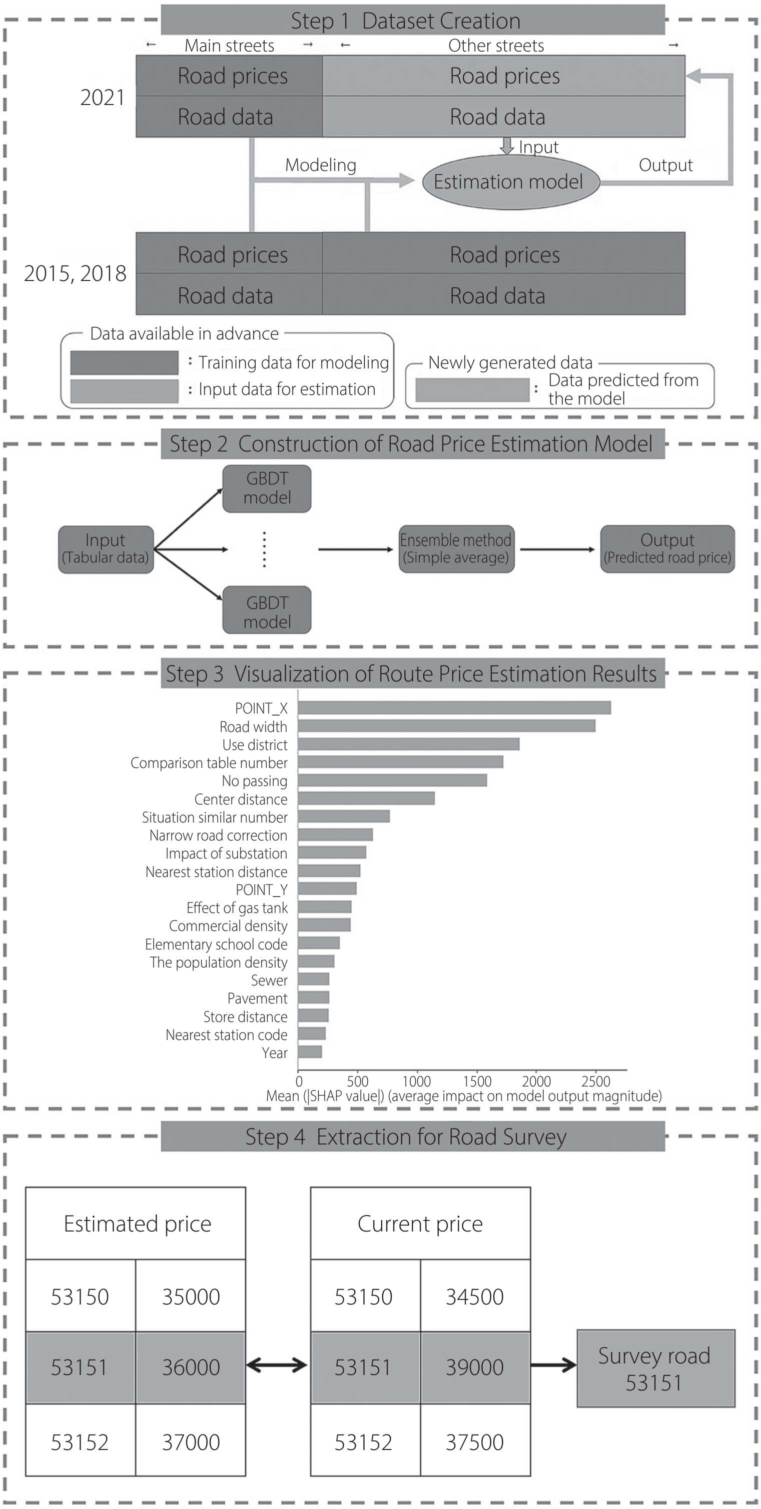
Feature	Explanation
Road type	National/prefectural/city roads, etc.
Road width	The road width (m)
Pavement code	Whether to pave the road surface with asphalt or concrete
No passing code	Whether or not vehicles can pass
Sidewalk code	Sidewalk on both sides/Sidewalk on one side/No sidewalk
One way code	Presence or absence of one-way traffic restriction
Pedestrian road code	Existence of pedestrian-only regulation
Narrow road correction code	Whether the width of the road satisfies the requirement of 2 m
Nearest station code	The nearest station
Nearest station distance	Distance to the nearest station (m)
Station name	Name of the nearest station
Store code	The nearest major commercial store
Store distance	Distance to the nearest major commercial store (m)
Elementary school code	The nearest elementary school
Elementary school distance	Distance to the nearest elementary school (m)
Central code	The nearest center area
Center distance	Distance to the nearest center area (m)
Sewerage code	Presence or absence of sewerage
Year	Year of road price data
Interchange code	The nearest interchange
Interchange distance	Distance to the nearest interchange (m)
Substation code	The nearest substation
Impact of substation	Distance to the substation
Gas tank code	The nearest gas tank
Effect of gas tank	Distance to the gas tank
Influence of JR Taketoyo Line	Distance from JR Taketoyo Line
Influence of Meitetsu Kowa Line	Distance from Meitetsu Kowa Line
Use district code	Use district by city planning
Specified floor area ratio code	Specified floor area ratio by city planning
Similar situation number	Further divided into smaller groups based on the situation of the use district
POINT_X	East–West coordinate of road measurement point
POINT_Y	North–South coordinate of route measurement point
Street type	Main street or other streets
Previous land code	Original land before land readjustment project was implemented
Comparison table number	Applicable comparison table number
Commercial density	Commercial utilization rate of buildings adjacent to the road
The population density	Average population of the town where the road is located
Transaction density	Number of transactions in the town where the road is located in the previous three years

using GBDT models, which are typically used for estimating tabular data. The model learns the training dataset constructed in step 1 and evaluates the test dataset to calculate road prices results. Finally, the ensemble method described in Section 5.1 is used to combine these models to improve the performance of the estimation model.

- In step 3, we use an evidence-based approach to visualize the importance of various features in determining road prices. This allows us to determine the formation factors for discussion and consideration. In addition, it allows us to objectively estimate the road prices and improve the transparency of road price determination.
- In step 4, we compare the estimated road prices with current road prices. Subsequently, we determine the roads that should be investigated further. Staff members can improve work efficiency by prioritizing surveys regarding estimated road prices.

5.3 Explainable AI for road price formulation

Attributes beyond accuracy are being increasingly empha-



Estimated price	
53150	35000
53151	36000
53152	37000

Current price	
53150	34500
53151	39000
53152	37500

Figure 8: Road price evaluation of other streets based on comparison table

sized owing to the popularization of AI utilization. In particular, fairness, accountability, and transparency are fundamental from a societal perspective. The "Human-Centered AI Social Principles" approved by the 2019 G20 specify principles of fairness, accountability, and transparency (Cabinet Secretariat, Japan, 2019). Explainable AI is essential in satisfying the three requirements of fairness, accountability, and transparency in AI. The explainability degree required depends on the area of AI utilization. In particular, the explainability of AI is essential in the administrative field. The "AI Utilization/Introduction Guidebook for Local Governments" formulated by the Ministry of Internal Affairs and Communications stipulates the principles of AI utilization (Ministry of Internal Affairs and Communications, Japan, 2022). For example, the verifiability of the input and output of the AI system, the possibility of AI services, and the interpretation of predicted results must be emphasized.

The SHapley Additive exPlanations (SHAP) method was used to calculate the contribution of each feature in order to visualize the relationship between each element (or feature) and the estimated price. The SHAP is a method deduced from game theory to explain the relationship between models and features, which allows the contribution of each feature to be understood more effectively (Lundberg and Lee, 2017; Lundberg et al., 2018; Lundberg et al., 2020). The SHAP calculates the contribution of individual players in game theory and assigns a SHAP value. The sum of the contributions of each feature matches the estimated value. For road price estimation, the SHAP value of each feature was calculated and the effect of each feature was visualized.

6. Experiment evaluation and analysis

In this section, the analysis and verification of the experimental results are provided.

6.1 Evaluation methods

Four methods are used to comprehensively evaluate and analyze the experimental results.

6.1.1 Mean absolute percentage error (MAPE)

The MAPE is the error between predicted and actual values expressed as a percentage, as shown in the formula below, where yi-pred is the i-th predicted road price, yi-true is the i-th actual road price, and n is the amount of data. The absolute value of the difference between the predicted and actual values divided by the actual value for each data point is calculated. Next, the average value obtained by dividing the total value by the amount of data is output. Generally, the average value is multiplied by 100 to obtain a number in percentage. The error can be the "predicted value - actual value" or the "actual value – predicted value."

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_{i\text{-}pred} - y_{i\text{-}true}}{y_{i\text{-}true}} \right| \tag{1}$$

6.1.2 Root mean squared error (RMSE)

The RMSE is calculated by averaging the squared error between the predicted and actual values for the data and calculating the square root of that value, as shown in the formula below, where yi-pred is the *i*-th predicted road price, $y_{i\text{-}true}$ the *i*-th actual road price, and n is the amount of data. If the calculation is based on the squared value of the error, then the RMSE is high when the data contain samples with significant errors.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{i\text{-}pred} - y_{i\text{-}true})^2} \tag{2}$$

6.1.3 Mean absolute error (MAE)

The MAE calculates the absolute error between the predicted and the actual values for the data, as shown in the formula below, where yi-pred is the *i*-th predicted road price, $y_{i\text{-}true}$ the *i*-th actual road price, and n is the amount of data. Because the MAE does not involve a squared error, it is less susceptible to outliers than the RMSE.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_{i\text{-}pred} - y_{i\text{-}true} \right| \tag{3}$$

6.1.4 Coefficient of determination (R^2)

The R^2 is an evaluation index that expresses how close the value predicted by a prediction model matches the actual value, as shown in the formula below, where $y_{i\text{-}pred}$ is the *i*-th predicted road price, $y_{i\text{-}aver}$ is the average value of the predicted values, $y_{i\text{-}true}$ is the i-th actual road price, and *n* is the amount of data.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n} (y_{i\text{-}pred} - y_{i\text{-}true})^2}{\sum_{i=1}^{n} (y_{i\text{-}aver} - y_{i\text{-}true})^2} \right) \tag{4}$$

6.2 Experimental results

The effects of using external features through were experimentally compared. In the experiments, in addition to the features of the comparison table used for estimating the road prices, we used the two external features mentioned earlier, i.e., the population density and transaction density, to calculate the road price.

Experiments were conducted using the four GBDT methods and ensemble methods described above (in Section 5.1). Table 2 shows the results of road prices estimation without using external features. When using a single GBDT model, such as the CatBoost model, the MAPE was be approximately 1.2607; the RMSE, approximately 1160.7380; the MAE indicator, approximately 557.1561; and the R^2, 0.9916. As shown in the results, the model achieved lower errors. Moreover, our proposed ensemble method for CatBoost, namely CatBoost Ensemble in the table, resulted in slight errors. For example, the MAPE was approximately 1.2607. Similarly, for other models, the results of the ensemble method were better than that of the single model. These results prove that even the simple ensemble method can improve the performance of road prices estimation and provide better results. Next, the ensemble models were combined using the single average ensemble method, which can improve road prices estimation. Consequently, the MAPE, RMSE, and MAE were 1.2242, 1146.1435, and 539.1555, respectively. The average road price was approximately 46,000. In contradistinction, when our model was used, the MAE was less than 540, which indicates that our model can reduce the average error of an estimated road price to an acceptable and practical level.

Table 3 shows the results of road prices estimation using external features. Compared with the experimental results shown in Table 2, the performance of each evaluation method deteriorated slightly after the external features were learned. The experimental results indicate that better performances can be achieved even when the external features are not learned. However, learning external features offers two benefits. First, the proposed model can learn and be tested also on the data of a specific city, such as Handa City. In the future, when this method is extended to other cities, better results can potentially be obtained after learning from external data from other cities. Second, learning from external data is an excellent example of applying open data to local government administrative tasks.

6.3 Analyses results

Figures 9 and 10 show the visualization results of the model without learning external features in terms of the SHAP value. Fig. 9 shows the importance of the features for the road prices. These are the features described in Table 1. Among them all, the road width exerted the most significant effect on the road price. Here, "road width" refers to the width of the road. According to the regulations, certain restrictions apply to constructions in local areas, depending on the road width. For example, a wider road allows vehicles to traverse it more conveniently. POINT_X is the second most important feature. In Handa City, the railway spans the north and south. The road prices at locations closer to the railway are higher, and lower at locations further away from the railway. Hence, the distance from the station in the East–West direction imposes a more significant effect.

Figure 10 shows the visualization results for a specific road. The width of this road is 17.2 m, significantly wider than the average road width of all roads, i.e., 5.7 m. Because the road spans a long distance, vehicles can traverse easily. A road with a large width exerts positive effects and results in high prices. However, the price has reduced owing to the geographical location of POINT_X. Additionally because the road is distant from the nearest station, the "nearest station distance" adversely affects the road price. Further to these factors, the use district and center distance adversely affect the road price. For example, the center is far from the nearest center, with its center distance being 4340 m, which confirms that this road is far from the city center. In general, the price of this road is low because of the lack of spatial location and the low degree of commercialization.

Figures 11 and 12 show the visualization results of the model when external features are learned, indicated in terms

Table 2: Results of road price estimation without learning external features

GBDT method	Evaluation methods			
	MAPE	RMSE	MAE	R^2
Xgboost	1.3193	1192.8775	581.0750	0.9911
Xgboost Ensemble	1.2840	1167.6018	569.5506	0.9915
Catboost	1.2607	1160.7380	557.1561	0.9916
Catboost Ensemble	1.2542	1155.8485	555.1516	0.9917
Ngboost	1.3146	1203.2785	592.5520	0.9910
Ngboost Ensemble	1.3010	1197.4233	574.5247	0.9910
LightGBM	1.2427	1153.5822	543.4877	0.9917
LightGBM Ensemble	1.2411	1154.6561	542.8840	0.9917
Ensemble method for all models	1.2242	1146.1435	539.1555	0.9918

Table 3: Results of road price estimation after learning external features

GBDT method	Evaluation methods			
	MAPE	RMSE	MAE	R^2
Xgboost	1.4463	1229.7886	640.3765	0.9906
Xgboost Ensemble	1.3584	1201.6723	606.4954	0.9910
Catboost	1.3998	1215.8490	621.8162	0.9908
Catboost Ensemble	1.3408	1184.7286	596.0093	0.9912
Ngboost	1.4336	1233.6536	632.1281	0.9905
Ngboost Ensemble	1.4074	1226.0240	624.0992	0.9906
LightGBM	1.3982	1195.4283	614.0505	0.9911
LightGBM Ensemble	1.3965	1194.3596	613.7074	0.9911
Ensemble method for all models	1.3241	1182.9506	590.3986	0.9913

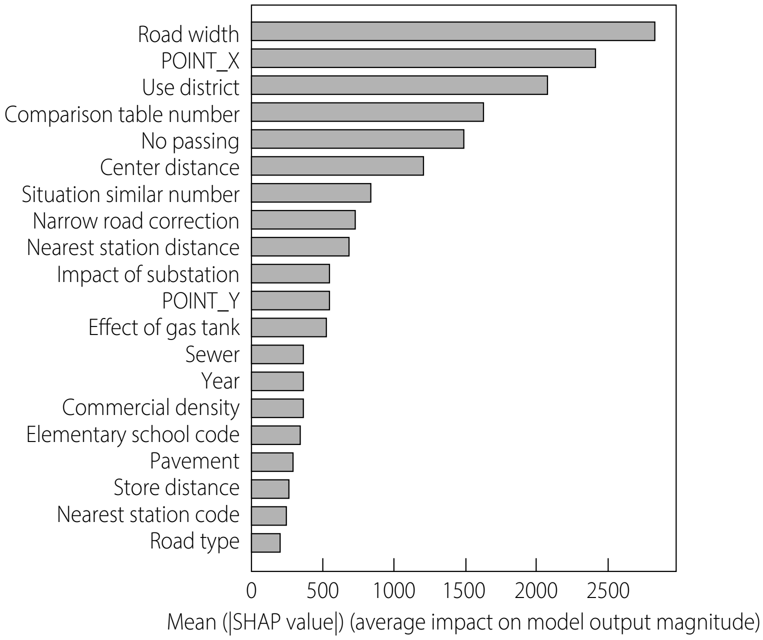


Figure 9: Importance of each feature of the estimation model without learning external features

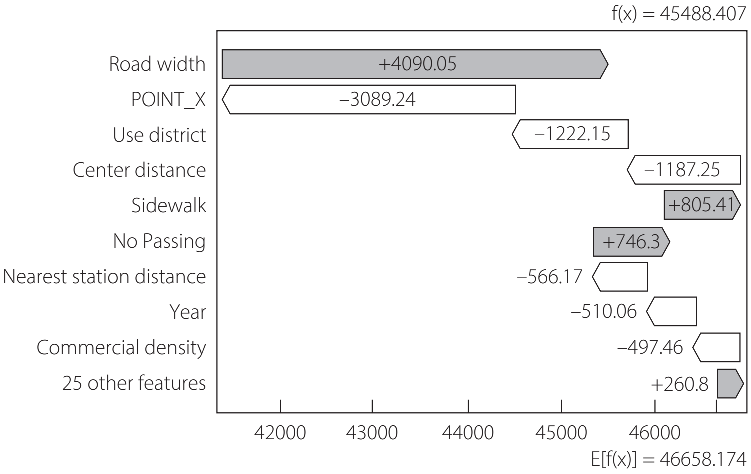


Figure 10: Visualization results for a specific road without learning external features

of their SHAP values. Fig. 11 shows the importance of each feature when the external features are learned in the road price estimation model. This indicates that the model does not significantly alter the prediction results when the learned features change slightly. From a comparison of the results shown in Figure 11 and 9, although the learned features are different, the trends of the overall features are similar. Meanwhile, the results shown in Figure 11 indicate that POINT_X and road width are the two most essential features, both with similar importance. On the basis of the degree of feature importance, the importance of the other features are similar in the two learning methods. Notably, the population density accounts for a certain degree of importance, which indicates that the population density affects road price estimation.

However, the estimation accuracy decreased after learning the features of population density and transaction density. This is attributable to the following three reasons: First, we calculated the population density based on the town where the road was located; the population density may not accurately reflect the actual situation of the corresponding road. Second, owing to insufficient data, we could not identify the town where the road was located, which made it difficult to determine the population density of the road. Third, transaction density is the number of land transactions divided by area in the past three years. Because land transaction data were not updated in real-time, time lags and data defects may occur, thus adversely affecting estimation. Figures 12 and 10 show the visualization results of the same road. A comparison of both visualization results shows that the estimated road prices and the relatively important features were

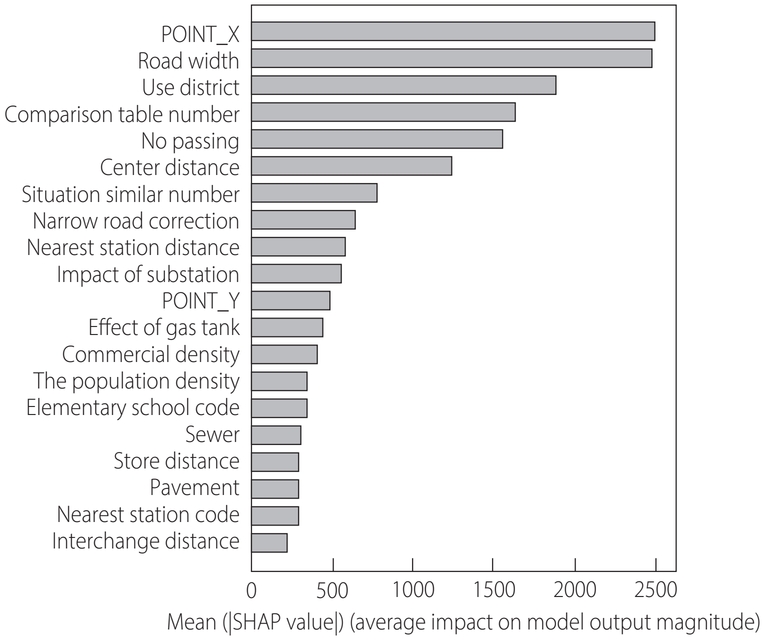


Figure 11: Importance of each feature of the estimation model when external features are learned

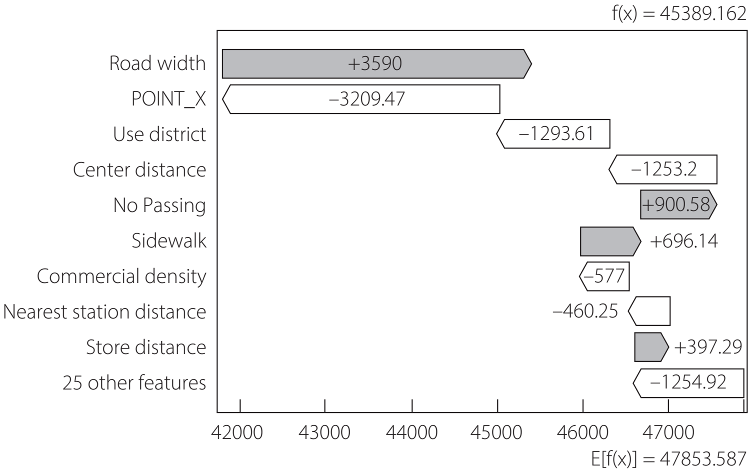


Figure 12: Visualization results for a specific road when external features are learned

similar. The order of importance of certain features might change slightly; however, the overall trend was similar.

In summary, we can comprehensively analyze and inspect the road price via visualization. The factors and reasons contributing to the setting of road prices can be understood easily even by non-experts.

6.4 Extraction of road data for priority survey

Our aim is to determine which road to prioritize surveys, increase work efficiency, and reduce workload. We compare the predicted road prices with those set by experts. If the predicted value differs significantly from the current value, the road to be surveyed was preferentially selected. For example, in the 2021 road data, experts adjusted the prices for 35 roads. The adjusted average road price was approximately 3774 yen, and the average error of our proposed model was approximately 664 yen. Our proposed model reduced the error by approximately 82 % compared with the price calculated directly using the comparison table. Among the 35 roads, the expert adjustment amount for a particular road was 6,400 yen, and the price difference between the price predicted by our proposed system and that estimated by the experts was only 24 yen, indicating that the proposed model can reduce road price estimation errors significantly, thereby reducing the work burden. In comparison, the expert adjustment amount for another road was 5,500 yen, and the price predicted by our proposed system was 4,008 yen, which is different from the prices predicted by the experts. Although our proposed system reduced the price difference to a certain extent, a relatively large error remained. Based on the magnitude of the price difference, we can prioritize the road with a more significant price difference for survey to increase work efficiency.

6.5 Utilization for new roads

Seven new roads have been constructed in Handa City, Aichi Prefecture, and the proposed system was used to predict the price of the new roads. The road prices could be directly predicted using the features of each road from the comparison table. Furthermore, visualization was performed to determine the features that affect each new road. These methods can be used as references when formulating the price of a new road. Moreover, the work efficiency and transparency could be improved. Using these two examples, the visualization results of the prices of new roads were explained below.

Figure 13 shows the visualization result of the first new road. Among several factors, road width exerted the most significant effect on the setting of the road price. The width of the road was 6 m, which is higher than the average road width and contributes positively to the road price decision. Moreover, in the "no passing" index, this road is passable for vehicles; therefore, it contributes positively to the price decision. Conversely, spatial location, POINT_X, center distance, and the nearest station distance all adversely affected the price decision. The center distance, which is the distance from the center to the nearest center area, was 3,360 m. The nearest station distance was 1,170 m, which may deter daily activities, thus decreasing the road price.

7. Conclusion

In this study, we discussed and analyzed the progress and existing problems of data utilization in Japan, particularly in the administrative field. Based on industry-government-academia collaboration, first, we proposed solutions and presented practical examples of data analyses, suggestions, and assistance in performing local government tasks. Second, we discussed the possibility of using open data as external data for model learning. Although this approach did not yield any positive effects, the findings obtained serve as a vital reference for the application of open data to the performance of local government tasks in the future. In the future, when the proposed model is expanded and applied to other cities, the imported external features may offer benefits. Third, we used the GBDT method to develop a road price formulation system based on existing data from public and private sectors. We analyzed the data possessed by local governments and proposed a system for automatically estimating road prices. Additionally, we effectively improved the estimation accuracy using an ensemble method. Fourth, model interpretability is crucial in applying AI, particularly in administrative tasks. By visualizing the results based on the SHAP value, the relationship between the predicted road price and each feature can be explained, thus allowing citizens to better understand the prediction basis as well as increase the credibility and interpretability of the system. Even a non-expert can understand the basis used to predict a road price. In the future, we will use more visualization methods to analyze the effects of features on road prices more comprehensively. Finally, our proposed system can be easily and quickly applied to new road data to predict road prices. Hence, the workload can be reduced, and the transparency and accountability in road price decisions can be increased.

Furthermore in the future, we plan to construct a comprehensive support system that utilizes machine learning (based on data from the public and private sectors) to evaluate fixed asset tax. Factors that are not listed in the comparison table but may affect the determination of roadside prices will be verified. Currently, experts are expected to fix road prices using the comparison table; however, they may consider factors that are not in the comparison table, such as the surrounding environment, when setting the prices. We wish to verify this and devise a method that can formulate road prices more effectively. The aim of this study was to improve the efficiency

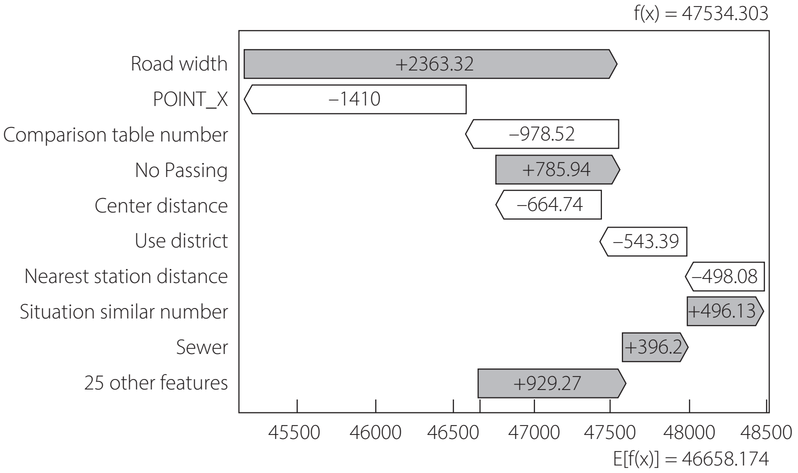


Figure 13: Visualization results for a new road without learning external features

and transparency of performance of local government administrative tasks. Lastly, in the future, a method utilizing ICT based on data analysis will be proposed to evaluate fixed assets such that data from public and private sectors (including fixed asset information in local governments nationwide) can be further utilized and developed.

Acknowledgements

We would like to express our gratitude to the Handa City General Affairs Department, Tax Division, for providing us with valuable data and useful advice. In particular, they provided us with the chance to demonstrate experiments. We are extremely grateful to all those concerned. This research was financially supported by JST SPRING, Grant Number JPMJSP2125.

References

Aoki, K., Takeda, K., Yano, K., and Nakaya, T. (2016). The study of proposed natural break method for land prices continuity to property tax. *Geographic Information System Association*, D-2-4. (in Japanese)

Cabinet Secretariat, Japan (2019). Social Principles of Human-Centric AI. Retrieved November 6, 2022 from https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.

Chen, T. and 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.

Chita Statistical Research Council, Japan (2020). Chita Peninsula Statistics. Retrieved November 6, 2022 from http://www.city.tokoname.aichi.jp/_res/projects/default_project/_page_/001/005/242/R2zenntaiban.pdf. (in Japanese)

Ding, J., Xue, N., Long, Y., Xia, G. S., and Lu, Q. (2019). Learning RoI transformer for oriented object detection in aerial images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2849-2858.

Ding, J., Xue, N., Xia, G. S., Bai, X., Yang, W., Yang, M. Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2021). Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 44, No. 11, 7778-7796.

Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., and Schuler, A. (2020). Ngboost: Natural gradient boosting for probabilistic prediction. *International Conference on Machine Learning*, 2690-2700.

Handa City, Japan (2022). Population and number of households in Handa city. Retrieved November 6, 2022 from https://www.city.handa.lg.jp/shimin/shise/toke/jinko/setai.html. (in Japanese)

IT Strategic Headquarters, Japan (2012). Open government data strategy. Retrieved November 6, 2022 from https://japan.kantei.go.jp/policy/it/20120704/text.pdf.

IT Strategic Headquarters, Japan (2016). Open Data 2.0. Retrieved Nov. 6, 2022 from https://cio.go.jp/sites/default/files/uploads/documents/opendata2.0.pdf. (in Japanese)

IT Strategic Headquarters, Japan (2017). Open data basic guidelines. Retrieved Nov. 6, 2022 from https://cio.go.jp/sites/default/files/uploads/documents/kihonsisin.pdf. (in Japanese)

JRI Review (2021). Science, are there enough local civil servants. Retrieved Nov. 6, 2022 from https://www.jri.co.jp/MediaLibrary/file/report/jrireview/pdf/12542.pdf. (in Japanese)

Kato, T., Endo, M., Urata, M. Yasuda, T., and Shimazaki, H. (2019). Creation of AI model for solar panel detection using aerial image and application to land evaluation. *The Society of Socio-Informatics*, SSICJ10-2. (in Japanese)

Kawano, Y., Endo, M., Urata, M. Yasuda, T., Shimazaki, H., and Kimura, T. (2020). Utilization of machine learning considering the distribution of predicted values for the verification work of property tax land prices. *The Society of Socio-Informatics*, SSICJ11-1. (in Japanese)

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, Vol. 30.

Ke, G., Xu, Z., Zhang, J., Bian, J., and Liu, T. Y. (2019). DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 384-394.

Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. (2017). Deep larning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, Vol. 14, No. 5, 778-782.

Li, Y., Aoki, K., Takeda, K., Irie, T., Furuichi, T., and Satoh, T. (2011). Applicability of kriging in the verification work of land tax assessment. *Geographic Information System Association*, D-3-5. (in Japanese)

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, Vol. 30.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K., Newman, S. F., Kim, J., and Lee, S. I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, Vol. 2, No. 10, 749-760.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence,* Vol. 2, No. 1, 56-67.

Ministry of Internal Affairs and Communications, Japan (2021a). FY2019 Settlement, White Paper on Local Public

Finance, 2021. Retrieved November 6, 2022 from https://www.soumu.go.jp/iken/zaisei/r03data/chihouzaisei_2021_en.pdf.

Ministry of Internal Affairs and Communications, Japan (2021b). Status of the number of local public servants. Retrieved November 6, 2022 from https://www.soumu.go.jp/iken/kazu.html. (in Japanese)

Ministry of Internal Affairs and Communications, Japan (2022). AI Utilization/Introduction Guidebook for Municipalities. Retrieved November 6, 2022 from https://www.soumu.go.jp/main_content/000820109.pdf. (in Japanese)

MLIT Japan (2022). Land General Information System. Retrieved November 6, 2022 from https://www.land.mlit.go.jp/webland/l. (in Japanese)

Prime Minister of Japan and His Cabinet, Japan (2016). Basic Act on the Advancement of Public and Private Sector Data Utilization. Retrieved Nov. 6, 2022 from https://japan.kantei.go.jp/policy/it/data_basicact/data_basicact.html. (in Japanese).

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, Vol. 31.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, Vol. 81, 84-90.

Takeda, K., Li, Y., Aoki, K., and Satoh, T. (2018). Investigation of route price verification method. *Geographic Information System Association*, D-1-4. (in Japanese)

The Government of Japan (2020). Declaration to Be the World's Most Advanced Digital Nation / Basic Plan for Promotion of Public and Private Sector Data Utilization. Retrieved November 6, 2022 from https://warp.ndl.go.jp/info:ndljp/pid/12187388/www.kantei.go.jp/jp/singi/it2/kettei/pdf/20200717/siryou1.pdf.

Tu, Y., Urata, M., Endo, M., Yasuda, T., Shimazaki H., and Kimura, T. (2021). Development of land use judgment system using deep learning to support land evaluation. Proceedings of *IEEE 10th Global Conference on Consumer Electronics*, 332-336.

Ukai, R., Endo, M., Urata, M., Yasuda, T., and Shimazaki, H. (2018). Use and application of aerial photographs and land category data using artificial intelligence. *The Society of Socio-Informatics*, SSICJ2017-1. (in Japanese)

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974-3983.

Yang, C., Rottensteiner, F., and Heipke, C. (2018). Classification of landcover and land use based on convolutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4*, Vol. 4, No. 3, 251-258.

Yu, J., Wang, Z., Majumdar, A., and Rajagopal, R. (2018). Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the United States. *Joule*, Vol. 2, No. 12, 2605-2617.

(Received: December 2, 2022; Accepted: December 24, 2022)