frontiers | Frontiers in Built Environment

TYPE Original Research
PUBLISHED 08 December 2025
DOI 10.3389/fbuil.2025.1631964

OPEN ACCESS

EDITED BY
Yakun Xie,
Southwest Jiaotong University, China

REVIEWED BY
Hourakhsh Ahmad Nia,
Alanya University, Türkiye
Zhixiang Xing,
Changzhou University, China

*CORRESPONDENCE
Yoshihiro Kabeyama,
s24d155@kagawa-u.ac.jp

RECEIVED 20 May 2025
REVISED 09 November 2025
ACCEPTED 17 November 2025
PUBLISHED 08 December 2025

CITATION
Kabeyama Y, Kajitani Y, Ueno T and Yuyama A (2025) Efficient disaster damage prediction method using building point data and LTSM: a case of flood disaster .
Front. Built Environ. 11:1631964.
doi: 10.3389/fbuil.2025.1631964

Efficient disaster damage prediction method using building point data and LTSM: a case of flood disaster

Yoshihiro Kabeyama[1]*, Yoshio Kajitani[1], Tsuyoshi Ueno[2] and Ayumi Yuyama[3]

[1]Faculty of Engineering and Design, Kagawa University, Takamatsu, Japan, [2]ENIC Division, Grid Innovation Research Laboratory, Central Research Institute of Electric Power Industry (CRIEPI), Yokosuka, Japan, [3]Structures and Earthquake Engineering Division, Sustainable System Research Laboratory, Central Research Institute of Electric Power Industry, Abiko, Japan

Accurate information on the location and use of individual buildings is essential for estimating impacts from disasters. However, even in developed countries, such data remains scarce, forcing reliance on aggregated statistics that obscure building-level impacts. We therefore propose a method for efficiently constructing point data on business facilities with industrial attributes for disaster analysis. We developed a multimodal industrial classification model within a Long Short-Term Memory (LSTM) framework. This model integrates business names from telephone directories with spatial context -business establishment statistics and land use zoning to probabilistically assign primary and secondary business types. As a result, an accuracy of approximately 83%–88% was achieved in industrial classification. The multimodal classification model contributed an average improvement of 13.0% in business establishment statistics and 5.4% in land use zoning for manufacturing predictions versus the non-multimodal case. The results of applying the damage and restoration functions from the manual to the prepared building data indicate variations ranging from 0% to 236% compared to a 500m grid-based damage method. The difference is significant compared to the accuracy of the building estimates, suggesting that it is desirable to change to building-based estimates.

KEYWORDS

flood disaster, building point data, industrial classification, LSTM, multimodal data fusion, natural language processing

1 Introduction

In recent years, intensifying torrential rains have caused large-scale flooding worldwide, and both their frequency and severity are projected to increase under climate change (Merz et al., 2010; IPCC, 2023). Similarly, Japan has experienced extensive damage to homes and businesses, as demonstrated by the 2018 Western Japan Heavy Rain Disaster and the 2019 Eastern Japan Heavy Rain Disaster. According to the Flood Damage Statistics Survey (Ministry of Land, Infrastructure, Transport and Tourism, 2018-2022a), both disasters affected numerous residential and commercial facilities nationwide, further

highlighting the importance of accurate flood damage estimation for advancing prevention measures and recovery planning.

Regarding flood damage estimation, it has been noted that a building's location and usage significantly influence the estimated damage losses (Fuchs et al., 2019; Dottori et al., 2016a; Merz et al., 2010). Recent research has led to the development of high-precision flood hazard maps (Oubennaceur et al., 2019; Dottori et al., 2016b; Bellos and Tsakiris, 2015; Hénonin et al., 2013). Moreover, building-specific damage loss models (Ma et al., 2024; Haque et al., 2023; Ran and Nedovic-Budic, 2016; Dias et al., 2018; Scorzini and Frank, 2017; Zabret et al., 2016; Zeng et al., 2019; Asgary et al., 2012; Kuroda et al., 2020; Ohara et al., 2022) are relatively accessible. Flood damage estimation models, such as flood depth damage models seen in prior research, have been enhanced by incorporating building attributes like number of stories, occupancy type, structural characteristics, and land conditions. This highlights the growing importance of identifying the attributes of buildings at risk of damage (Chen et al., 2025; Paulik et al., 2023; Englhardt et al., 2019; Scorzini and Frank, 2017). Darnkachatarn and Kajitani (2025) demonstrated differences in vulnerability by industrial sector in Bangkok through surveys, contributing to improved accuracy in estimating business asset losses. However, detailed building databases essential for properly utilizing such economic loss models are lacking. For large-scale estimates, it is common to allocate urban assets aggregated by administrative districts or grids. This leaves a gap between high-resolution hazards and building-level databases where business activities are systematically coded. The objective of this study is to construct a building database retaining business type information by establishment, focusing on flood damage models.

Japan can utilize statistical data surveyed by the national government as individual record data as a building database suitable for exposure in flood risk assessment models. However, accessing this data requires numerous processes like applications and reviews, making its use generally difficult and burdensome in terms of time and cost. Furthermore, business statistics created through large-scale face-to-face surveys are updated infrequently, making them unsuitable for estimating annual losses. Therefore, it is necessary to develop database construction methods that use more readily available data. Internationally, many studies have evaluated natural disaster risks using OpenStreetMap (OSM) (Cerri et al., 2021; Ullah et al., 2023). However, as open data, OSM faces challenges regarding data update frequency and the absence of basic features in some regions, affecting its reliability (Goldblatt et al., 2020). Tu et al. (2023) evaluated flood risk in Shanghai using Baidu Maps, a Chinese map search service. They estimated damage costs by utilizing the recorded building shapes, number of floors, and land use, comparing damage costs by affected area and building use. While such readily available building databases exist in some regions, their use across all areas is difficult. Therefore, Bhuyan et al. (2022) constructed a database by using OpenStreetMap and Google Maps to assist in building type classification. Further, they detected building shapes from satellite and aerial imagery using the ResU-Net model, a type of convolutional neural network (CNN). A key challenge with this methodology is the significant time required for the image training process and classifying individual building occupancy types, making it difficult to generalize a consistent workflow from training data preparation. Additionally, research has shown that using land use data as open data improves damage estimation by building (Esparza et al., 2025). Furthermore, a design-based approach for impervious surface extraction using VHR SAR × GLCM, etc., is an effective option for urban asset identification that can complement the spatial bias of OSM (Polverino et al., 2024). While the above approaches enable building database construction to capture building distribution and broad usage, there remains no established methodology for building databases with detailed industry classifications suitable for advanced economic damage modeling.

Furthermore, regarding research estimating building usage from business establishment databases, Akiyama and Shibasaki (2011) proposed evaluating name similarity using NLP based on facility names and addresses, improving data integration accuracy from approximately 20%–85%. Tojo and Oyama (2022) estimated usage from names using TextCNN, achieving approximately 80% accuracy. However, this was a 5-class classification, and concerns remain about the limitations of CNN's local feature learning in multi-class scenarios. Meanwhile, Maki et al. (2023) analyzed telephone directory names using LSTMs, achieving 85% accuracy across 22 industries. These results indicate that industry classification from names is approaching practical levels. However, the industry classifications from these studies are limited and insufficient for flood damage estimation, a challenge which this research aims to address.

Therefore, the core of this study is to propose a method for efficiently constructing a business establishment database. This involves using publicly available building databases supplemented by telephone directory data to assign detailed industry classifications. Classification prediction is performed through simple matching of telephone directory data and text analysis using LSTMs. Furthermore, recognizing that land use data is effective for estimating building purposes, we construct an integrated multimodal model that also uses spatial information as supplementary data. To the best of the authors' knowledge, no previous study has applied multimodal classification combining LSTM and spatial information to industrial forecasting. This research contributes to the field by quantitatively evaluating the effectiveness of business statistics and land use zoning in multimodal classification for improving industry prediction results, demonstrating the potential utility of spatial information for industrial classification forecasting. Furthermore, this study estimates the error range between damage estimates based on historical cases using this database and estimates using the conventional 500 m grid as a 100% benchmark, and compares this with the impact of multimodal classification on industrial prediction errors. Thus, this study emphasizes the importance of

Abbreviations: MLIT, Ministry of Land, Infrastructure, Transport and Tourism; GSI, Geospatial Information Authority of Japan; MIC, Ministry of Internal Affairs and Communications; CNN, Convolutional Neural Network; RNN, Recurrent Neural Network; LSTM, Long Short-Term Memory; NLP, Natural Language Processing; API, Application Programming Interface; "main", Indicates the primary industry type in the industry estimation results; "sub1", Indicates the first secondary industry type in the industry estimation results; "sub2", Indicates the second secondary industry type in the industry estimation results; "max1", Indicates the probability of being classified as the "main" in the results of industry estimation; "max2", Indicates the probability of being classified as the "sub1" in the results of industry estimation; "max3", Indicates the probability of being classified as the "sub2" in the results of industry estimation.

utilizing detailed point-level building data. The workflow developed in this study enables the construction of a building database containing industry information with significantly less effort than interview surveys, and allows for annual data updates. This is expected to reduce the loss gap in past flood damage estimation and future loss prediction compared to grid-based estimation, thereby appropriately supporting flood damage estimation.

This study is structured as follows: Section 2 describes the building database workflow and target area, Section 3 compares the base data, Section 4 presents the results of multimodal classification using LSTM, Section 5 provides verification through damage cost estimation and external comparison, Section 6 discusses the findings, and Section 7 concludes.

2 Materials and methods

2.1 Method for constructing a building database

This study aimed to construct a detailed building database enabling appropriate damage estimation from publicly available datasets. Specifically, the goal was to build a database compatible with the asset damage evaluation indicators in the Ministry of Land, Infrastructure, Transport and Tourism's (2019) "Flood Countermeasure Economic Survey Manual (MLIT 2018–2022)". This manual outlines damage assessment methods by building use (residential, business premises), and for business premises damage, it specifies asset damage evaluation values by major industry category under the Japan Standard Industrial Classification. For the manufacturing sector, since loss evaluation amounts vary significantly with more detailed classifications, individual evaluation amounts are provided. Therefore, this study develops a business establishment database containing industry information aligned with the manual. The number of target classes was organized into 23 classes (17 non-manufacturing classes and 6 manufacturing classes). Class labels are shown in Supplementary Table A2.

The building data that forms the foundation for database construction is the building point data provided by Zenrin Co., Ltd. (2018), which includes building names and high-precision coordinate information. This data also contains information such as the number of stories and total floor area, which significantly influence flood damage estimation. Analysis of building location conditions before and after flooding by Ito et al. (2019) demonstrates the effectiveness of this data for flood damage assessment. Furthermore, to examine the validity of using the building point data, Chapter 3 compares the number of households and businesses in statistical data. The building point data is enhanced by adding industry information to the building names. For the business establishment list containing industry information, electronic telephone directory data (Japan Software Service, 2022) with extensive industry information was adopted. First, the simplest method for adding industry information is matching the building names in the building point data with the company names and address data in the telephone directory. While data matched using this method is highly reliable as industry information, the matching conditions are susceptible to variations in notation between the datasets, making it difficult to combine large amounts of data. Therefore, a method for assigning industries that does not depend on the notation of building names is necessary.

Recent advances in text-based data classification have led to the development of various techniques (Li et al., 2022; Zhu and Cao, 2024; Kim, 2014). Industry classification based on building names requires accurate category prediction from short text data. Short sentence classification is generally considered challenging due to difficulties in feature space representation and word association. However, Dos Santos and Gatti (2014) and Lee and Derenoncourt (2016) demonstrated the effectiveness of neural networks for this task. Regarding applications to land use analysis, the CNN model by Tojo and Oyama (2022) and the LSTM model by Maki et al. (2023) suggest that the LSTM model may be capable of predicting business types from establishment names with higher accuracy across more classes. This is thought to be partly because establishment names often exhibit sequence-dependent collocations, such as brand names, industry descriptors, and branch/location terms. Convolutional models excel at detecting local n-grams through convolution and pooling, performing well when industry information is conveyed via short, consecutive patterns like suffixes or affixes in names. In contrast, LSTMs effectively capture such sequential and positional dependencies more effectively than bag-of-words features or shallow feedforward baselines (Kim, 2014; Dos Santos and Gatti, 2014; Lee and Derenoncourt, 2016). As such sequential dependency patterns frequently occur in this research task, the sequential induction bias of LSTMs aligned better with the data generation process. However, challenges persist when predicting business industry classifications, including industry-specific training data imbalances and the substantial computational cost of LSTMs, leaving text-based classification with limitations in both accuracy and efficiency.

Therefore, to develop a deep learning model enabling more accurate industry classification, we explored the use of data sources related to business establishments. Bhuyan et al. (2022) achieved an F1 score of 74% for estimating building usage by employing cluster analysis based on building shape and supplementing it with open data such as OSM, Google Maps, and land use data. They thus demonstrated that spatial information can improve the classification accuracy of building usage. Furthermore, Teo et al. (2025) and Neffke et al. (2011) point out that industrial location and geographic attributes show a strong correlation. This study examines two spatial information datasets: the number of establishments by industry from business statistics and land use zoning. The Establishment Statistics provide nationwide boundary information for establishment locations and reflect the actual distribution of establishments by industry. Therefore, it is considered to play a supplementary role in guiding prediction results toward aligning with actual conditions. Conversely, land use data consists of boundary information for limited areas designated as urban planning zones. Compared to the Establishment Statistics, its overall effect is considered smaller. However, as this information pertains to zones that systematically regulate the location conditions for each building use, it is considered useful for adjusting prediction results toward specific industries that meet the regulatory conditions within those zones. The zoning information for land use data is shown in Supplementary Table A3. These boundary data exhibit trends similar to the distribution of facilities by building use and are

considered highly useful for building use classification prediction. However, there is a concern that using them simultaneously may cause overfitting due to redundancy. This study experimentally clarifies whether these data can function as effective data for industry classification based on diversity classification.

Therefore, we applied a model trained on the relationship between company names and industries in the telephone directory, targeting the building names in the building point data. Furthermore, we constructed a multimodal industry classification model. This model employs LSTM, which is more effective at capturing sequence dependencies in business names than non-time-series baselines, and is supported by location-dependent clues from spatial feature fusion. As additional justification for adopting LSTM, we compared four text classification models (Transformer, DNN, TextCNN, and LSTM). The results are shown in the Supplementary Material.

As a supplement, a key consideration when using telephone directories as training data is that they are collected assuming daily consumer services and sales. This means specialized industries and small-to-medium-sized businesses may be omitted, while, conversely, duplicates may occur, such as multiple department numbers within the same business. Business duplication can often be resolved through name or address consolidation. However, the potential for overfitting due to uneven distribution of business counts across target industries remains a challenge. Furthermore, the industry classifications used in the telephone directory employ unique names distinct from the Japanese Standard Industrial Classification (JSTIC). Consequently, the classification destination may not always align with business statistics. Here, the industry classifications in the telephone directory are pre-mapped to JSTIC before being used as training data. This correspondence is shown in Supplementary Table A2.

2.2 Analytical framework

In this study, we developed a building database that classified residential and business buildings based on the framework shown in Figure 1 and used it to estimate flood damage. The estimation accuracy was verified by comparing the predicted damage with actual flood damage data. Residential and business buildings were classified according to the building classification included in the building point data. For residential buildings, the number of households was adjusted using vacancy rates from the Housing and Land Statistics (Ministry of Internal Affairs and Communications, 2018) and evaluated using the coefficient of determination (R^2) by comparing the estimated distribution with regional population statistics (Ministry of Internal Affairs and Communications, 2015). For business buildings, manufacturing and non-manufacturing establishments were classified using the developed industrial classification model. Their distribution was evaluated using R^2 by comparing the results with regional business establishment statistics (Ministry of Internal Affairs and Communications, 2014).

The industry classification model consisted of three processes: (1) extracting business buildings based on building type classification in the building point data, (2) matching them with telephone directory data, and (3) performing text analysis using LSTM. This model was applied to the building point data to classify building usage patterns efficiently and systematically. For text analysis using LSTM, a multimodal model was constructed by integrating an NLP model based on the telephone directory data with additional factors, such as business distribution ratios within a 500 m grid (Ministry of Internal Affairs and Communications, 2016) and land use zoning classifications from urban planning regulations (Ministry of Land, Infrastructure, Transport and Tourism, 2019). The classification accuracy was evaluated by testing different data combinations to optimize the model.

To validate the damage estimation, we selected 11 regions with a history of flooding and integrated the building databases with inundation estimation maps from previous large-scale floods. (Geospatial Information Authority of Japan, 2018). The result of damage estimation was performed following the Flood Control Economic Manual, and the results were compared with conventional grid-based estimates to highlight the advantages of using point data. Furthermore, the accuracy of flood damage estimation was verified by cross-referencing industry classifications and flood impacts with a business damage survey in selected areas (Nihei et al., 2020).

The correspondence between the study areas, flood occurrence years, and the period of the building point data used is summarized in Figure 2 (Hereafter, area names are replaced with the symbols shown in Figure 2). Table 1 illustrates the damage situation in these areas. This study further utilized business statistical data from the Economic Census for Business Frame (Ministry of Internal Affairs and Communications, 2014) and the Economic Census for Business Activity (Ministry of Internal Affairs and Communications, 2016). The former emphasizes fundamental attributes of businesses and provides detailed industry classifications, while the latter focuses on economic activities, such as sales, and includes only broad industry categories. Consequently, the 2014 dataset was used for comparison with building point data and employee number estimation, while the 2016 dataset was used for training the industry classification model.

3 Comparison of building point data and statistical data

3.1 Basic information and aggregation method of building point data

This study utilized building point data (Zenrin Co., Ltd., 2018) as the foundational dataset for the residential and business building database. This dataset included location information and attribute data, such as building name, building area, number of floors, and building classification. Additionally, the data can be accessed at the city, ward, town, or village level, with buildings in the target area provided as individual data points with coordinates. This structure facilitated the aggregation of building attributes across various boundary units.

Moreover, the building point data included buildings categorized into residential (1,000 series), business (2000 series), commercial facility (3,000 series), and others (9,999). To ensure dataset accuracy, we verified whether the number of households and business establishments in the area aligned with

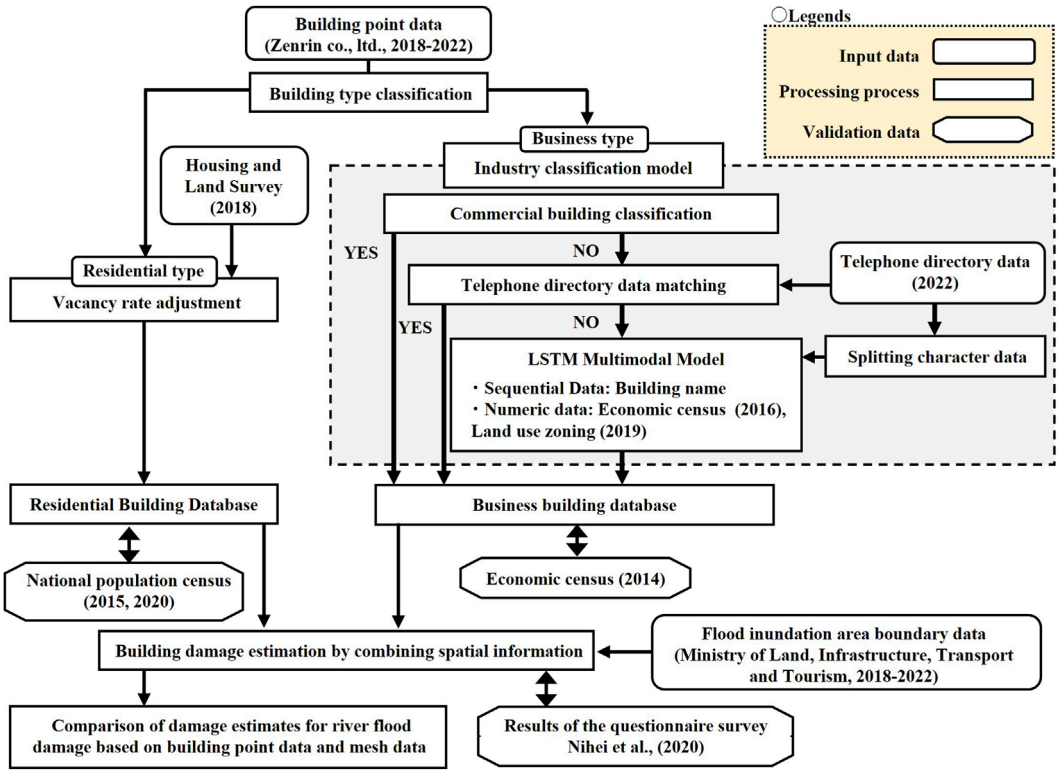


FIGURE 1
Industry classification framework for building point data.

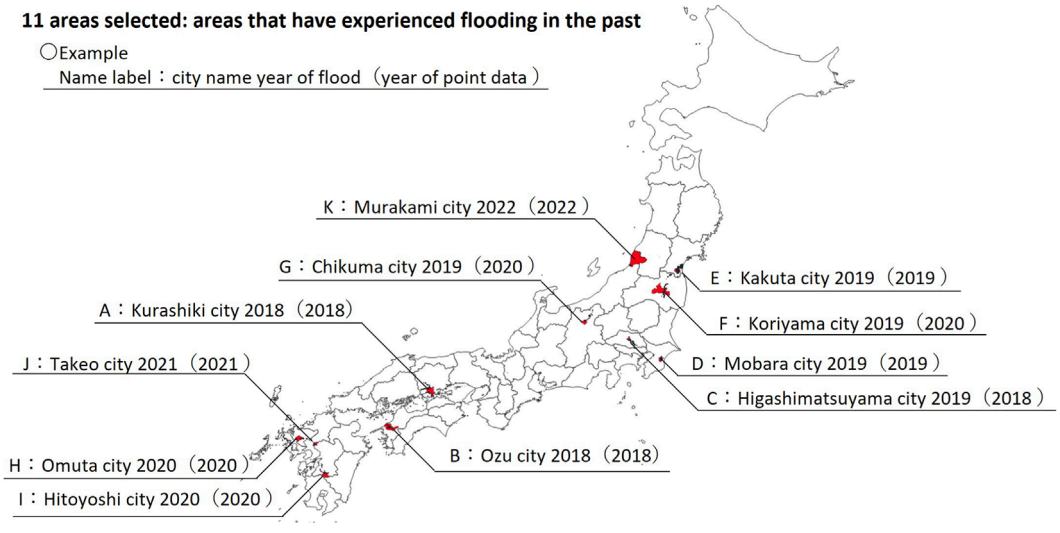


FIGURE 2
11 areas selected that have experienced flooding in the past.

statistical data. For comparison, we used population statistics (Ministry of Internal Affairs and Communications, 2015) for the number of households and business establishment statistics (Ministry of Internal Affairs and Communications, 2014) for the number of business establishments. These statistical datasets were selected as the closest in date to each flood occurrence. The method for compiling the building point data is outlined in Table 2. The number of households and businesses was collected for each subdivision of administrative districts in each region to ensure consistency in aggregation units with the statistical data.

TABLE 1 List of flood damage in 11 selected areas.

City label	Year of disaster	Number of houses damaged by the disaster (buildings)	Number of affected business establishments (buildings)	Number of employees affected by the disaster (persons)	General assets and operating suspension losses (million yen)
A	2018	5,858	642	5,157	239,738
B	2018	3,961	968	6,776	79,408
C	2019	580	42	716	18,624
D	2019	3,967	580	225	58,978
E	2019	2,487	274	1,137	26,139
F	2019	5,069	477	22,115	168,493
G	2019	828	33	743	15,013
H	2020	634	246	245	17,404
I	2020	4,871	65	6	126,148
J	2021	1,791	275	1,568	37,581
K	2022	1,561	141	511	13,827

TABLE 2 Building type and aggregation method.

Class	Building type	Building type number	Aggregation unit	Counting method
Residential building	Detached house	1001,1008	Number of households	Total number of private homes
	Apartment complex	1,002–1,006		
Office building	Business establishment	2001–2027	Number of establishments (including some households)	Total number of establishments
Commercial building	Shopping malls, tenant buildings, etc. (including some residential buildings)	3,001–3,004		
Others	Vacant houses, vacant businesses, vacant landetc.	9,999	Disaggregation	Disaggregation

3.2 Estimating the number of households

The total number of households based on the census (Ministry of Internal Affairs and Communications, 2015) and building point data for the 11 regions was compared using administrative division units for aggregation and evaluated by R^2, an index showing the goodness-of-fit of the estimated values (Figure 3). The evaluation results ranged from 0.656 to 0.979 for all the regions, confirming that the estimation error was small in most regions and consistent with the demographic data. However, in some regions, the number of households in the building point data was underestimated. This is because the building point data is created using visual confirmation and municipal data, making it easy to determine whether a detached house is occupied, yet difficult to verify the occupancy status of apartment buildings. Several buildings classified as "apartment buildings" in the building point data have zero individual units, which may indicate that the buildings are vacant. The tendency for the estimated number of households to be underestimated suggests that the occupancy status of apartment buildings is not understood.

We calculated the vacancy rate of apartment buildings by region using the Housing and Land Statistics (Ministry of Internal Affairs and Communications, 2018) and compared it with the point data. The vacancies for apartment buildings based on the point data were 25%–63% higher. To improve the accuracy of the household estimates, we corrected the number of private households in apartment buildings in the building point data. The correction method involved estimating the number of households in apartment buildings with zero private units, aligning the estimate with the

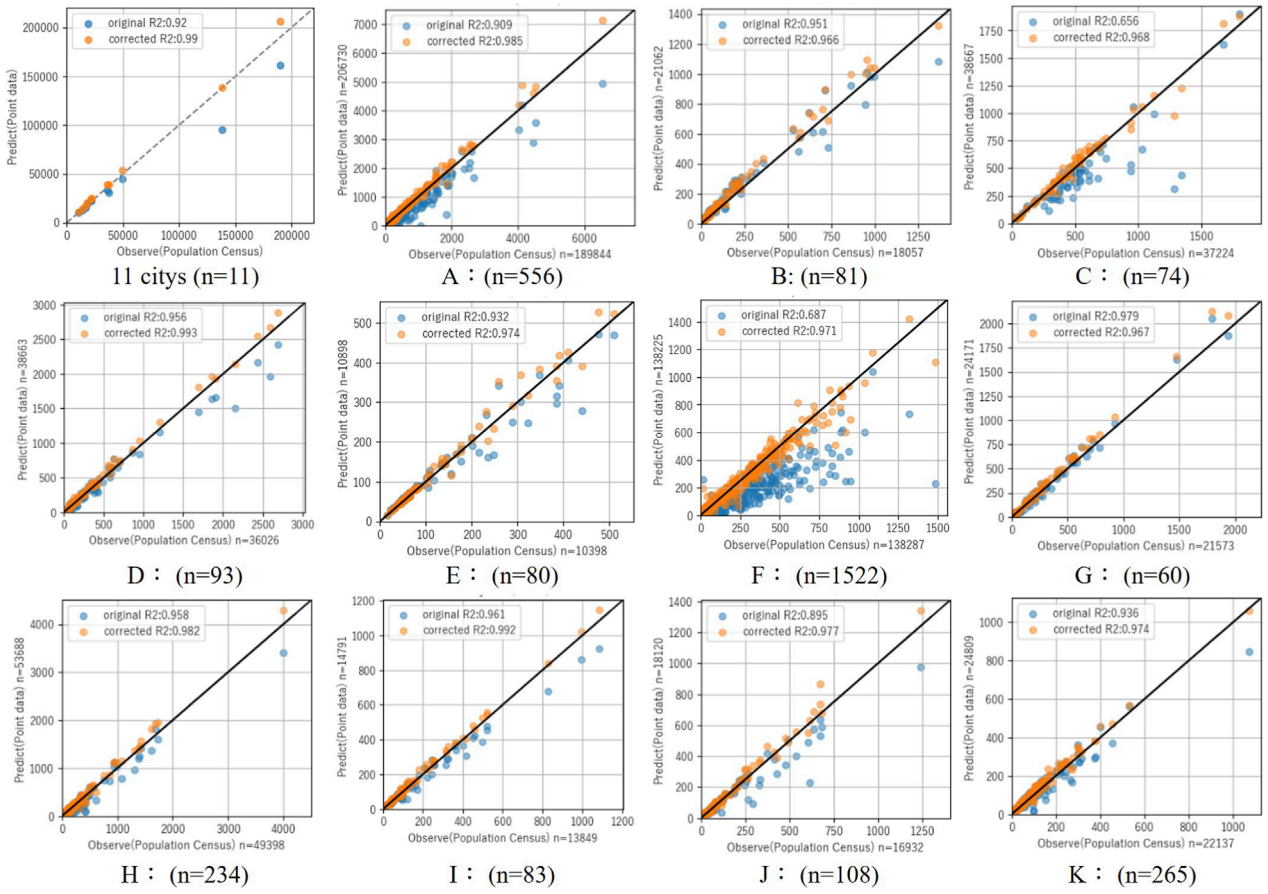


FIGURE 3
Comparison of household numbers.

vacancy rate of apartment buildings in the corresponding region based on the Housing and Land Statistics.

The vacancy rate of apartment buildings by region was calculated using the Housing and Land Statistics (Ministry of Internal Affairs and Communications, 2018) and compared with the corresponding rate derived from the building point data. The vacancy rate based on the point data was 25%–63% higher. To improve the accuracy of household estimates, buildings with existing household data were assumed to be based on surveyed data. Thus, they were retained without modification. In contrast, the number of households in buildings with zero private units was estimated based on regional vacancies. The total number of households, combining both retained and estimated values, was adjusted to match the vacancy rates for apartment buildings in each region as reported in the Housing and Land Statistics.

As a result of the correction, R^2 values increased to between 0.957 and 0.989 in all the regions. In Region C, where the fit improved the most, R^2 increased from 0.656 before correction to 0.957 after correction. This suggested that it was feasible to accurately estimate the number of households using building point data.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (y_i: \text{Observed}, \hat{y}_i: \text{Predicted}, \bar{y}: \text{Residual Mean})$$

3.3 Estimating the number of business establishments

Figure 4 shows the results of comparing the number of establishments based on the establishment statistics (Ministry of Internal Affairs and Communications, 2014) and building point data, aggregated by administrative division, for the 11 regions. The R^2 values were evaluated to be high, ranging from 0.800 to 0.995 for all the regions, indicating that the establishment distribution was accurately captured.

When aggregating the number of establishments based on building point data, it is important to note that since the point data digitizes each building individually, some establishment data may involve multiple buildings, such as warehouses, on the same premises. Therefore, integration processing is required when comparing with statistical data. In this process, buildings with the same name and address are treated as a single establishment. However, some establishment data may not be fully integrated due to inconsistencies in the notation of names. Nonetheless, the error is minor in all the regions, and the impact on the estimation of establishment damage is limited. Overall, the building point data showed a consistent distribution of establishment locations in alignment with the establishment statistical data.

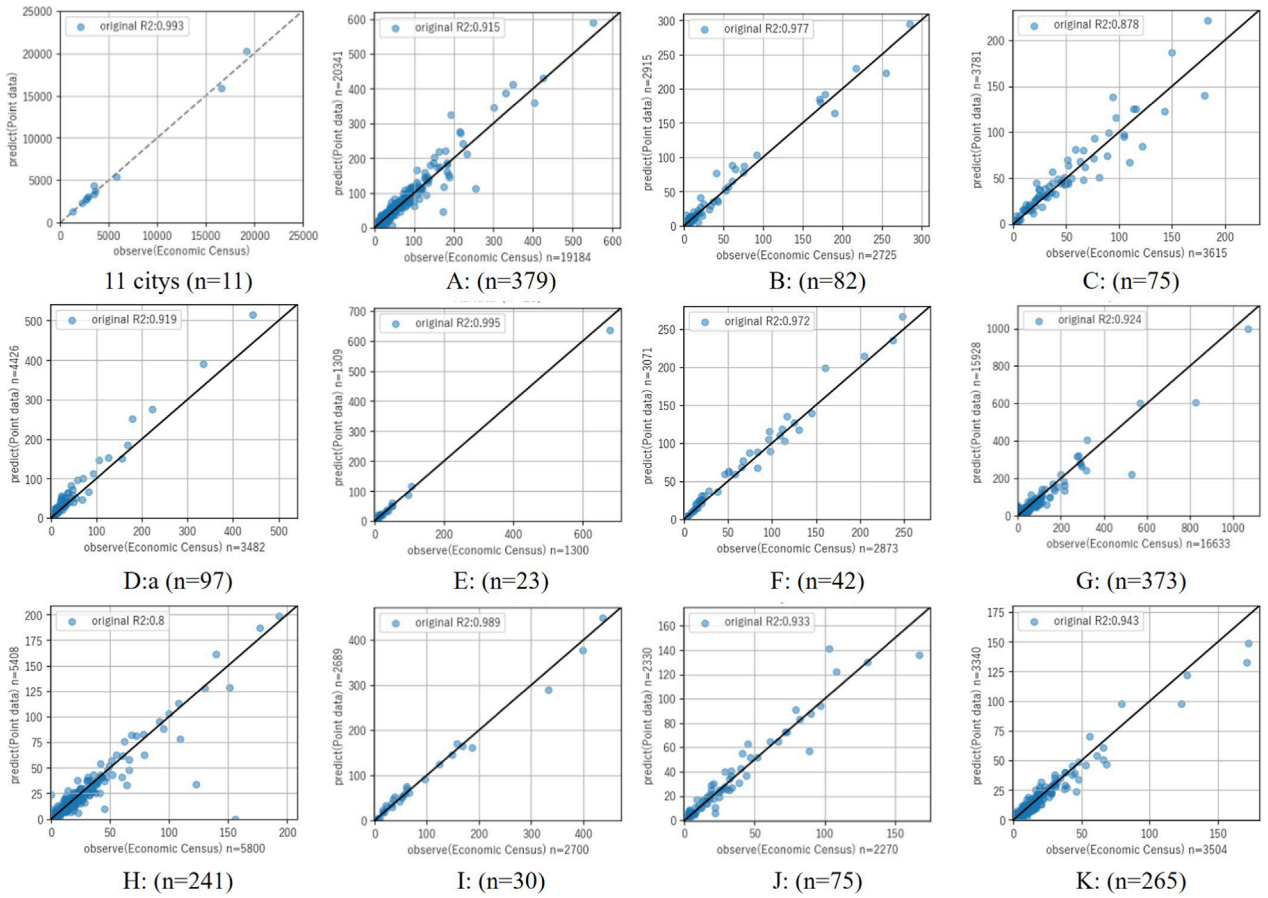


FIGURE 4
Comparing observed and predicted number of establishments.

4 Construction of the industry classification model

4.1 How to construct a multimodal industry classification model using LSTM

To classify business establishment data into industries, we constructed a multimodal model that integrates text analysis using LSTM with spatial information. The layer structure of the deep learning model is shown in Figure 5.

For the text analysis using LSTM, the company names and their corresponding industries from the telephone directory data (Nippon Software Service, 2022) were used as training data. The classification involved 23 categories, including 17 classes from the major industry classifications based on the Japan Standard Industrial Classification (Ministry of Internal Affairs and Communications, 2023), excluding manufacturing, and six additional classes representing similar business forms derived from the manufacturing medium classification. The industry classification model was built using Keras, a deep learning library for Python.

Since the company names in the telephone directory data are in Japanese, morphological analysis was required as a pre-processing step. The Python library, Janome, was used to perform morphological analysis, breaking down company names into common words and proper nouns. These decomposed words were converted into numerical data using Keras' keras. Preprocessing.text function, which served as the initial input for the industry classification model. Additionally, to implement a multimodal model based on building name analysis, we incorporated spatial information. This included the ratio of establishments by industry within a 500 m grid data from the Establishment Statistics (Ministry of Internal Affairs and Communications, 2016) and the land use zoning based on the City Planning Act (Ministry of Land, Infrastructure, Transport and Tourism, 2019). These data points were used as input to capture the spatial context of each establishment's location.

To integrate this spatial information with the telephone directory data, we obtained the coordinates for each establishment by searching their addresses in the telephone directory using the GSI Application Programming Interface (API).

4.2 Evaluating industry classification models

Model training was conducted on 3,549,019 business establishment data items, with duplicate company names removed

Layer structure of LSTM model

Layer (type)	Output Shape	Param #
input (Input Layer)	(None, 15)	0
embedding (Embedding)	(None, 15, 128)	329186048
dense (Dense)	(None, 15, 128)	16512
lstm (LSTM)	(None, 64)	49408
dense (Dense)	(None, 32)	2080
dense (Dense)	(None, 23)	759

Layer structure of LSTM multimodal model

Layer (type)	Output Shape	Param #
input (Input Layer)	(None, 15)	0
embedding (Embedding)	(None, 15, 128)	329186048
dense (Dense)	(None, 15, 128)	16512
input (Input Layer)	(None, 16)	0
lstm (LSTM)	(None, 64)	49408
dense (Dense)	(None, 32)	2080
concatenate (Concatenate)	(None, X)	X
dense (Dense)	(None, 32)	3104
dense (Dense)	(None, 23)	759

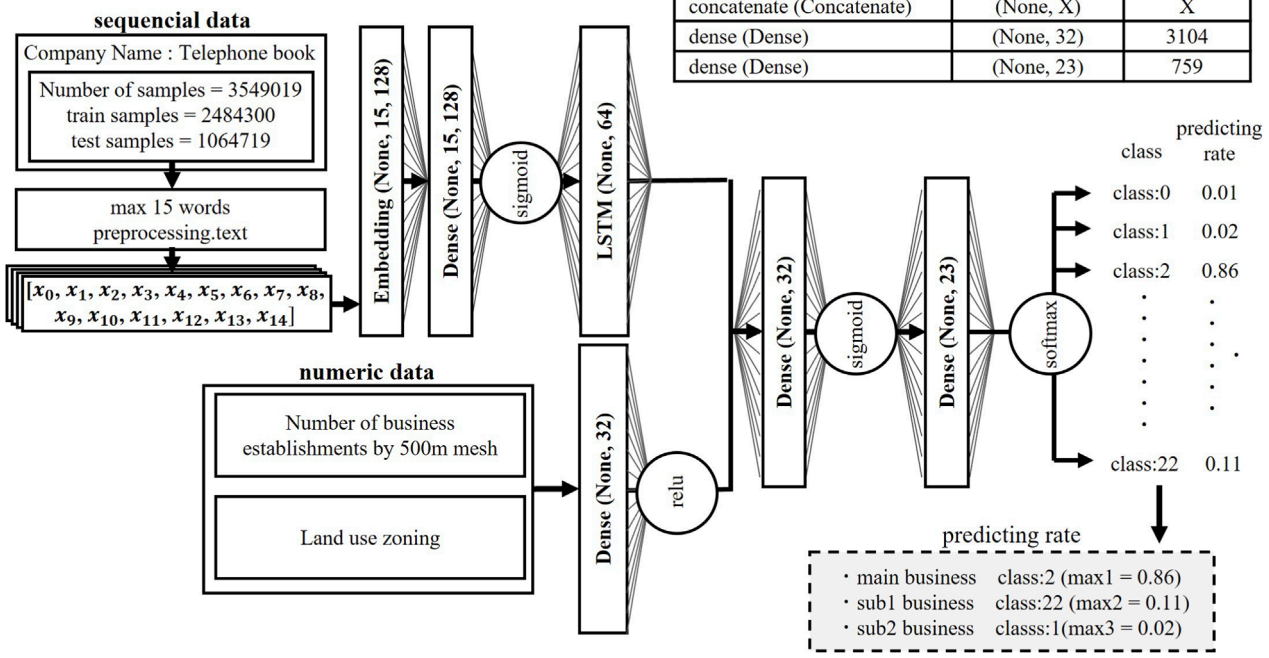


FIGURE 5
Overview of the LSTM-based multimodal models for industry classification.

from the telephone directory data. In this case, the sampling method warrants consideration. However, Maki et al. (2023) found that using nationwide data yielded approximately 82% accuracy in precision comparisons based on training data sampling methods, whereas training with regional data resulted in approximately 78% accuracy. Therefore, we adopted a configuration that uses nationwide data uniformly. The data was divided into training and validation sets, with 70% allocated for training and 30% for validation. The data was classified into 23 industry categories. During classification, the softmax function was used to predict the probability of each data item belonging to each class. This function returned the probability $P_{(y_i=1)}$ for each class. In this study, the class with the highest predicted probability ("max1") was treated as the primary industry type ("main"), the second-highest probability ("max2") as the first secondary industry type ("sub1"), and the third-highest probability ("max3") as the second secondary industry type ("sub2").

$$P_{(y_i=1)} = \frac{e^{x_i}}{\sum_{k=0}^{n} e^{x_i}} \quad (i = 0, 1, , , n) \quad n = 22$$

Figure 6 shows the results of applying the classification model to the validation data from the telephone directory. The graph displays industry classes on the x-axis, and the results include the overall accuracy rate for all the data (0–22), including non-manufacturing (0–16) and manufacturing (17–22). The bar graphs for each class represent the output results for the following four models (from left to right): ① LSTM, ② LSTM + Business establishment ratio (500 m grid), ③ LSTM + The land use zoning, and ④ LSTM + Business establishment ratio (500 m grid) + The land use zoning.

As a result of the classification in the learning process, although there was no significant difference in classification accuracy across models, the multimodal model showed improved accuracy, particularly in manufacturing and certain other industries. Model ④, which had the highest classification accuracy, achieved a 72% rate for the "main," 11% for "sub1," and approximately 5% for "sub2" across all industries. When considering additional industries, the accuracy improved to 83%–88%. This suggested that certain words in company names may be related to multiple similar industry types, and the results demonstrated that the classification accuracy could be effectively enhanced by creating a model that predicted the primary and the additional industries. Regarding the impact of business establishment statistics and land use zoning on industry forecasts, the multimodal model integrating statistical and zoning data showed a clear improvement in certain low-sample industries. Nonetheless, the overall improvement in average accuracy was

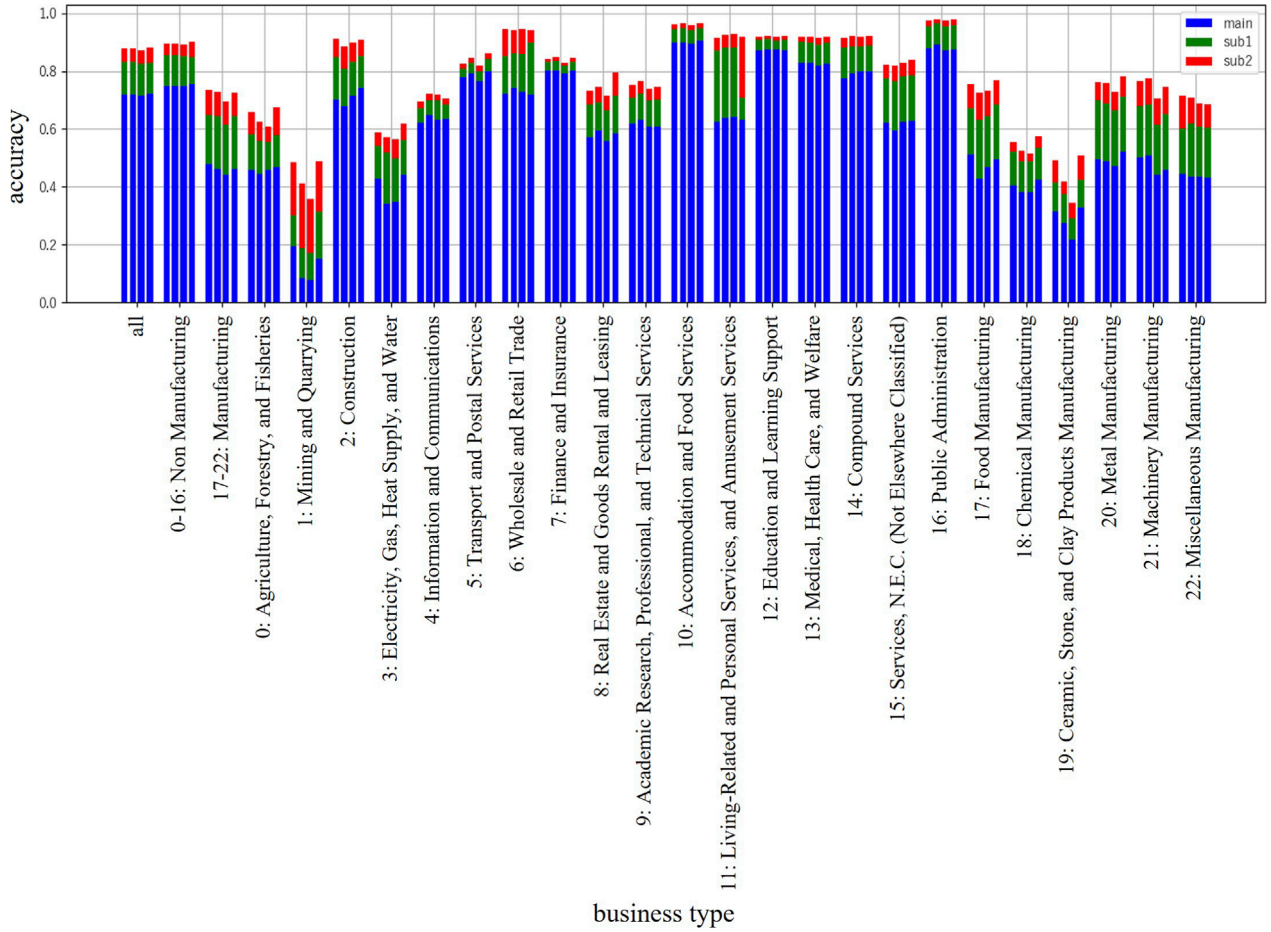


FIGURE 6
Comparison of classification accuracy of telephone directory data.

limited compared to the telephone directory data alone. In particular, the average improvement rate turned positive for industry groups with sample ratios of less than 2%, suggesting that latent industry distribution characteristics that cannot be expressed by textual information alone are complemented by statistical and geographic variables. Conversely, for the major industries with high sample ratios, the contribution of the additional variables was limited or appeared as a slight performance degradation because the phone book data already contained sufficient information. This suggests that multimodal integration is a structure that contributes more to the correction of data-sparse industries than to overall accuracy. The impact of the model's industry classification error on flood damage estimation is examined in the discussion in Chapter 6.

4.3 Classification method for building point data

To achieve efficient and highly accurate industry classification for building point data, a rational industry classification process was conducted before applying the LSTM text analysis model. The first step focused on processing commercial facility building data. Identifying the industry from the name of a commercial building is often challenging. Since multiple business establishments can exist within a single building, classifying the industry based solely on the building name is difficult as well. Therefore, the number of business establishments by industry within the building point data was tallied, and the primary and subordinate industries were determined based on the ratio. However, some data lacked the number of business establishments by industry, and these entries were moved to the next processing stage for further handling.

Next, a building name matching was conducted using telephone directory data for the same address. Specifically, based on the listed company name and address, building names in the building point data that matched the addresses of the administrative divisions were combined and assigned industry data. The number of business establishments identified through these processes and the total number of data points requiring industry prediction by the LSTM model are shown in Table 3. Approximately 30%–40% of the data could be classified during pre-processing, highlighting the importance of developing a model capable of efficiently predicting industry classifications.

10.3389/fbuil.2025.1631964

TABLE 3 Number of matching results.

Commercial building classification	A	B	C	D	E	F	G	H	I	J	K
Industry classification process	1,353	150	205	250	61	1,290	173	317	163	118	154
Telephone directory data matching	5,621	846	1,109	1,325	440	4,470	971	1823	684	863	1,309
LSTM	14,605	2,786	2,755	3,336	1,195	10,841	2,490	3,997	1724	2,170	2,652
Total	21,579	3,782	4,069	4,911	1,696	16,601	3,634	6,137	2,571	3,151	4,115

4.4 Industry classification accuracy in Higashimatsuyama City (C)

To confirm the accuracy of industry classification for individual business establishments against building point data, we used the building point data of Higashimatsuyama City (C) and the corporate database (2024) owned by Tokyo Shoko Research (2024). This was used to perform a matching process of industry data based on data, such as building names and addresses. Figure 7 shows the results of comparing the prediction results of the industry classification model with 777 data whose industries were identified as correct answers. Model ① had the highest accuracy rate for all industries, with an accuracy rate of 0.73% when the correct answer was up to the secondary industry. This result is a lower evaluation than the classification accuracy using the verification data of the telephone directory. In particular, the classification accuracy of the manufacturing industry was low. It was suggested that accurately predicting a single industry as the correct answer for building point data may be difficult. However, in the Tokyo Shoko Research database, the headquarters industry of the company is matched as the industry; therefore, it may not match the industry of the actual business establishment. This may have caused the accuracy rate to decrease in this analysis.

4.5 Comparison of business classification results of building point data and business establishment statistical data

The trained industry classification model using LSTM was applied to 11 regions. The predicted industry classes were aggregated into two categories–non-manufacturing (0–16) and manufacturing (17–22). The industry location distribution, as predicted by the model, was evaluated by comparing it with establishment statistics (Ministry of Internal Affairs and Communications, 2014) for each administrative district subdivision. The classification models were compared using the four models described previously, and the aggregation methods were performed in three patterns: (1) primary industry only, (2) primary industry + secondary industry, and (3) primary industry + secondary industry + tertiary industry. The calculation methods for the share of each industry in the sample are presented in Equations 1–3. The aggregation results are shown in Figure 8.

$$\text{"main"} : main_{rate} = 1.0 \tag{1}$$

$$\text{"main"}+\text{"sub1"} : main_{rate} = \frac{max1}{max1+max2}, sub1_{rate} = \frac{max1}{max1+max2} \tag{2}$$

$$\text{"main"}+\text{"sub1"}+\text{"sub2"} : main_{rate} = \frac{max1}{max1+max2+max3}, \quad sub1_{rate} = \frac{max2}{max1+max2+max3}, \quad sub2_{rate} = \frac{max3}{max1+max2+max3} \tag{3}$$

As a result, in the non-manufacturing industry, the accuracy rate was high across all regions, with no significant differences in the output results for each model. However, R^2 improved when using a calculation method that considered the additional business type. This result is consistent with the outcomes for the telephone directory data, suggesting that similar performance can be achieved for building names in building point data. In the manufacturing industry, some regions showed a significant improvement in the R^2 value when using the multimodal model, indicating that incorporating building names and spatial information was effective for classifying industries within building point data. The percentages of improvement by multimodal classification are shown in Table 4, and the values shown are the difference in the variation of scores when model (1) is set to 1.0. On average for all regions, improvement is possible by 13.0% for (2), 5.4% for (3), and 15.4% for (4). In particular, there is an extreme improvement in region D due to the number of establishments by boundary, and improvements are also observed in regions B, H, I, and J. The scatter plots plotted at the regional level show that the number of establishments predicted to be in the manufacturing sector decreases and approximates the published value in the other models relative to the results predicted in (1). In other words, the spatial data used here seem to work strongly in the direction of eliminating establishments that were predicted to be in the manufacturing industry by LSTM text classification based on their names. However, they were judged to have a high probability of not being in the manufacturing industry based on spatial information. In addition, the regions with strong improvement are considered to have fewer manufacturing establishments and a higher concentration of manufacturing establishments as a regional characteristic. Additionally, it was confirmed that accuracy improved when considering the occupation type, similar to the non-manufacturing industry. Hence, when estimating the distribution of industries from business establishments in a region, it is effective to use prediction probabilities that account for the occupation type, rather than relying solely on the primary industry classification.

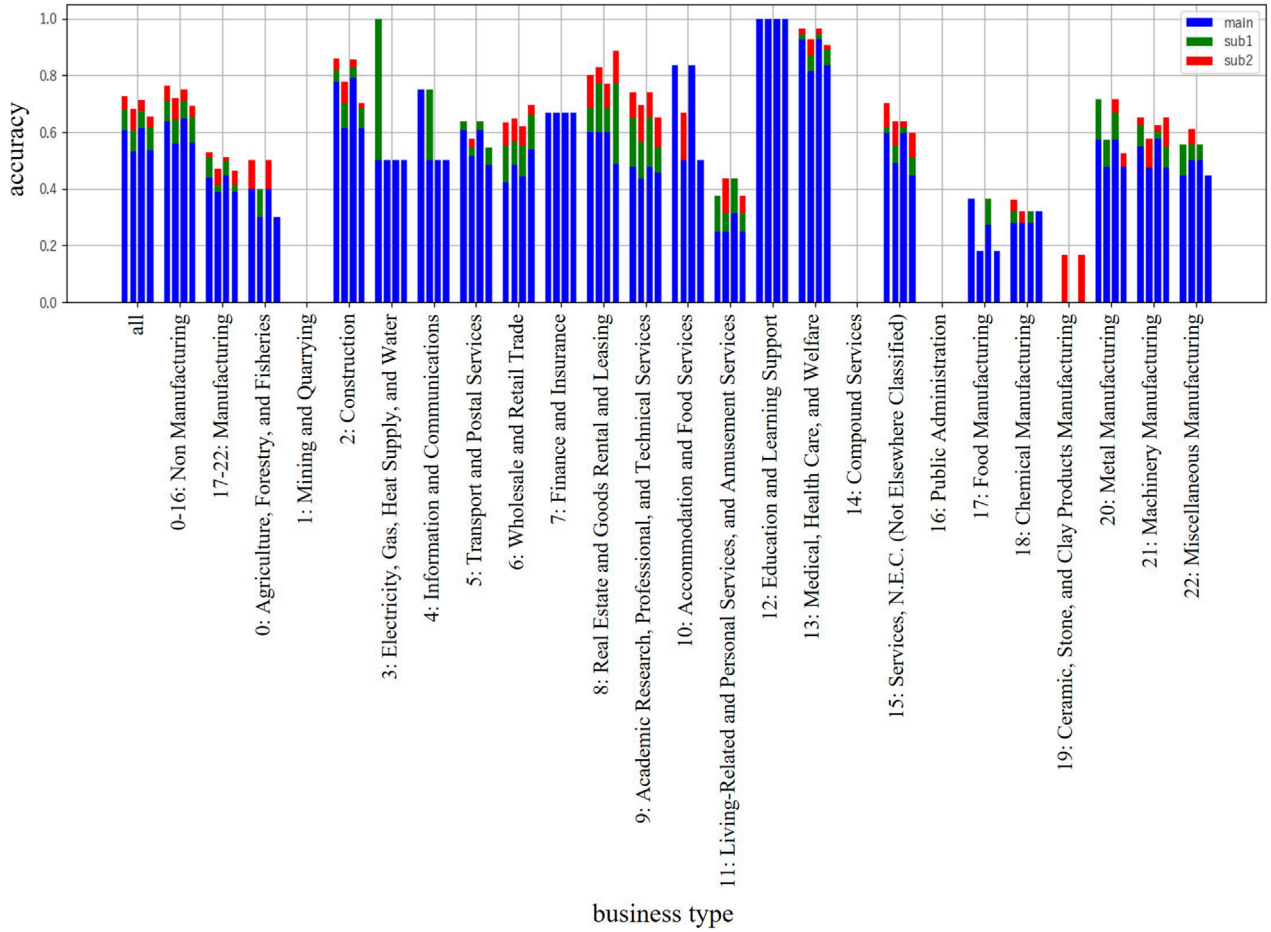


FIGURE 7
Industry classification accuracy for Higashimatsuyama City.

5 Implications for flood damage estimation

5.1 Comparison of damage estimates for flood damage based on building database and grid data

To estimate flood damage using building point data, we compared it with the inundation estimation map (Ministry of Land, Infrastructure, Transport and Tourism, 2018–2022b) published by the Geospatial Information Authority of Japan and calculated the damage using the estimation method outlined in the Manual for Flood Control Economics (Ministry of Internal Affairs and Communications, 2018). For residential damage, the amount of damage to household goods was estimated based on the number of affected households. For business damage, the total damage to depreciable assets and inventory was estimated based on the number of employees in each building. The number of employees used for business damage estimation was derived by aggregating the number of employees by industry in the administrative district where the business building was located and apportioning this based on the total floor area of the building. Damage estimates for each building were evaluated by integrating building coordinates and flood depth maps. The software utilized Python's geopandas, with the coordinate system set to WGS84.

Additionally, as a benchmark, we performed an estimation using the 500 m grid from business statistics (Ministry of Internal Affairs and Communications, 2014) and compared the estimates based on building point data with the grid-type estimates. The damage estimate using business statistics was calculated by considering the number of employees by industry in the grid, the average inundation depth, and the ratio of the inundated area to the total area of the 500 m grid. The parameters used in the damage estimation method using a 500 m grid are identical to those used in the estimation method using building point data. Residential damage is the aggregated number of households, and business damage is the aggregated number of employees affected. Figure 9 presents the results of comparing the estimated damage amounts by region. The trends in residential and business damage were consistent across regions. When the grid-based estimate was taken as 100%, the damage amounts estimated from point data ranged from 19% to 143% of the grid-based estimate.

Figure 10 shows the relationship between damage amounts by grid. When comparing damage by grid, the change in the

10.3389/fbuil.2025.1631964

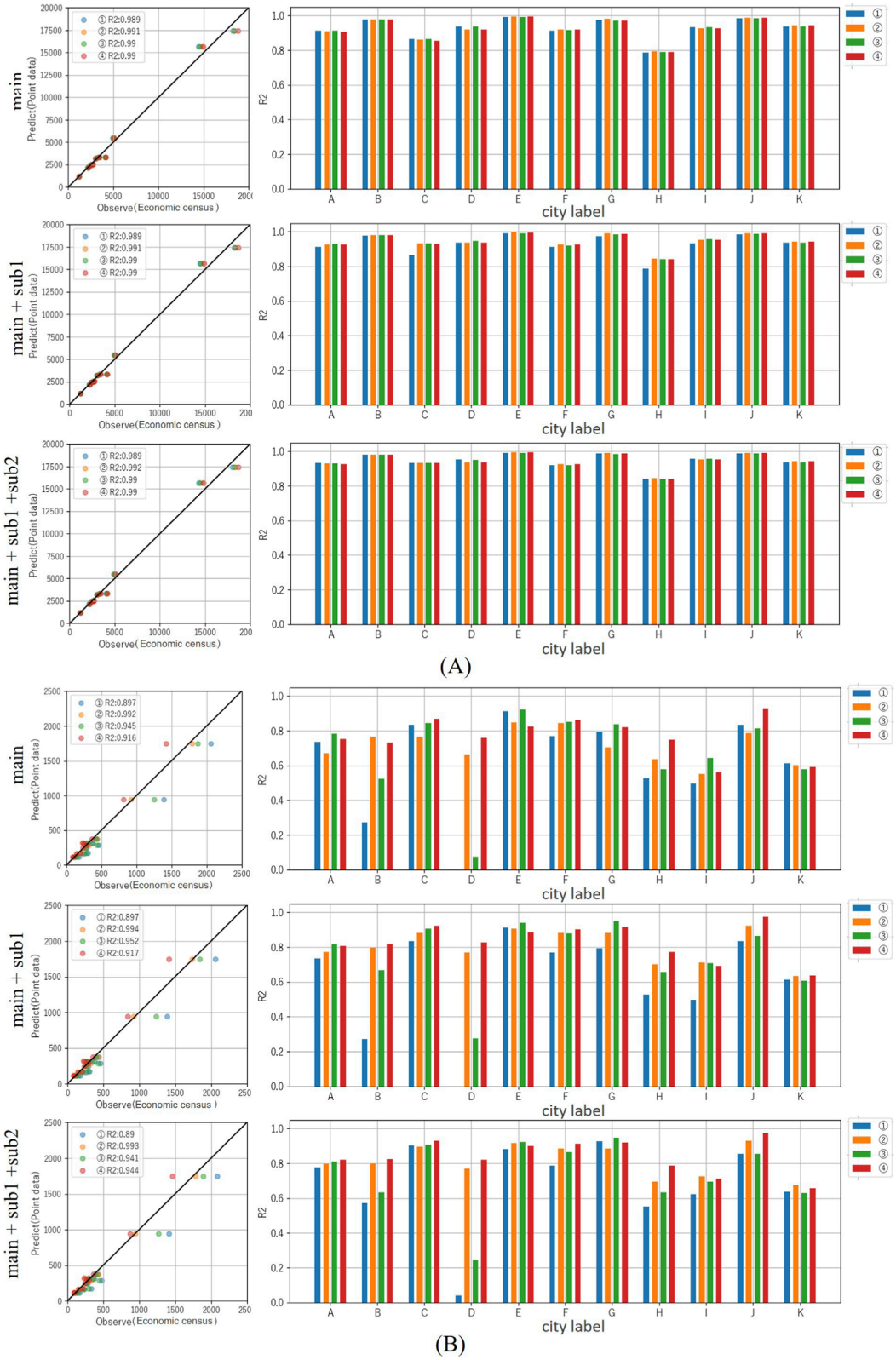


FIGURE 8
Evaluations of the goodness-of-fit of location distribution based on the industry classification results of each model for the region. **(A)** by non-manufacturing industries **(B)** manufacturing industries.

TABLE 4 Degree of improvement in the variation in the number of manufacturing establishments relative to model (1) for main + sub1 + sub2 by 11 regions.

Regions	R^2 score of ①	Improvement rate			Number of manufacturers per 500 m grids		Industrial area/area of the whole region (%)
		②	③	④	Average	Standard deviation	
A	0.777	0.021	0.033	0.045	1.65	2.35	0.073
B	0.572	0.229	0.06	0.251	0.50	0.98	0.000
C	0.901	−0.006	0.005	0.028	1.35	2.37	0.006
D	0.039	0.731	0.204	0.781	0.53	0.97	0.000
E	0.882	0.035	0.042	0.017	0.53	1.10	0.014
F	0.785	0.1	0.081	0.126	0.94	1.80	0.010
G	0.926	−0.041	0.021	−0.008	1.88	2.11	0.000
H	0.55	0.145	0.084	0.235	1.17	1.86	0.122
I	0.624	0.101	0.072	0.086	0.77	1.25	0.000
J	0.856	0.075	0	0.118	0.80	1.34	0.000
K	0.637	0.037	−0.008	0.02	0.71	1.17	0.000
ALL	0.686	0.130	0.054	0.154	—	—	—

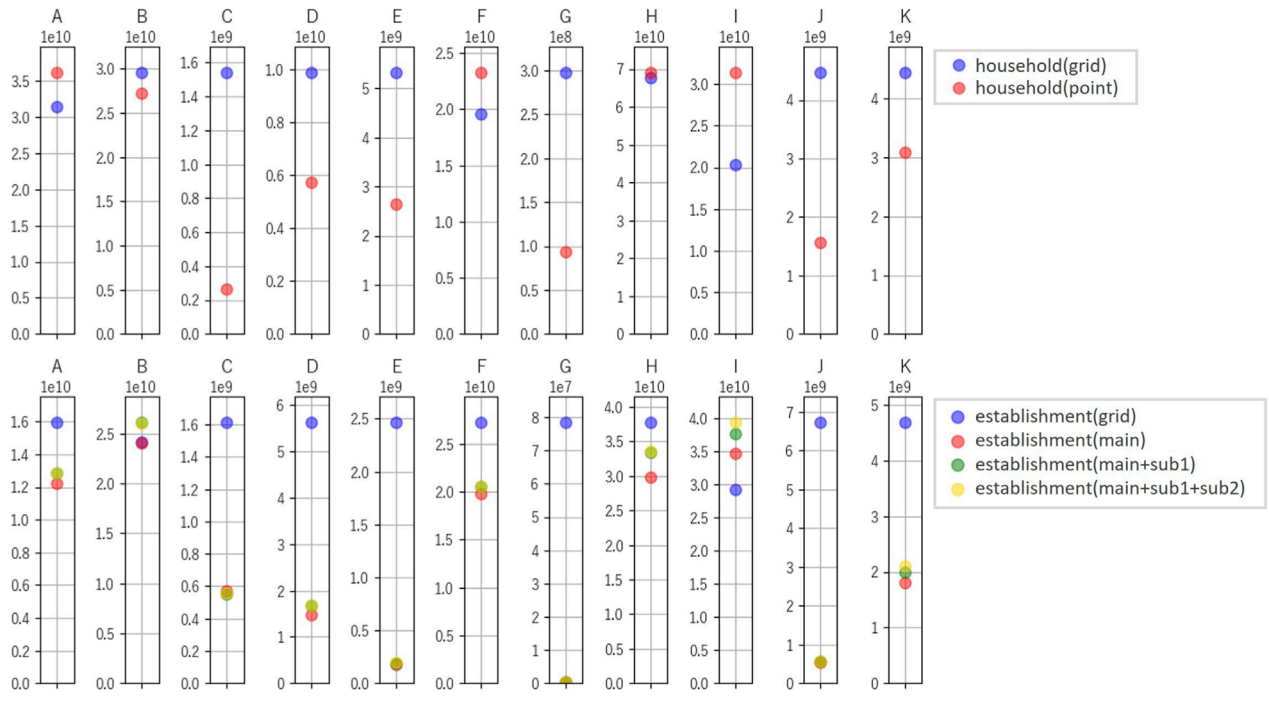
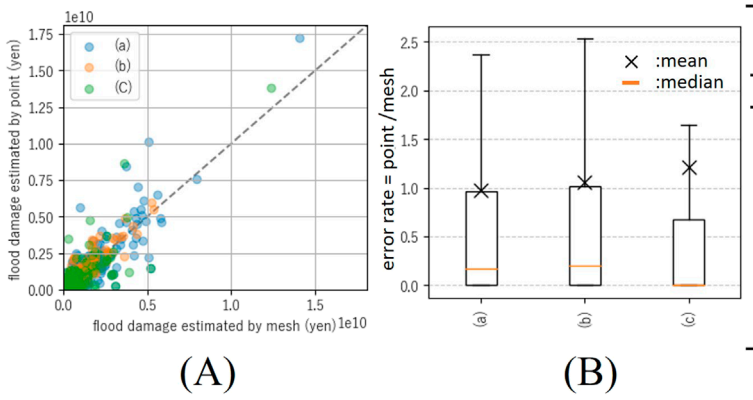


FIGURE 9
Comparison of damage estimates by city using mesh data and point data.

estimated damage amounts was significant, ranging from 0% to 236%, excluding outliers (those falling outside the range of 1.5 times the interquartile range). This indicated that grid-based estimates could overestimate or underestimate damages. Overestimation occurred because grid-based estimates assumed average damage for households and businesses that were not affected. This result is



		(a) all data	(b) household	(c) establishment
Overrated samples		1067	1037	829
Error rate	mean	0.978	1.057	1.214
	std	2.916	3.260	5.540
	min	0.000	0.000	0.000
	25%	0.000	0.000	0.000
	50%	0.173	0.199	0.000
	75%	0.960	1.021	0.672
	max	54.662	54.662	106.259

(C)

FIGURE 10
Comparison of damage estimates by 500 m mesh using mesh data and point data. **(A)** Error plot for each mesh **(B)** Boxplot of error bars (excluding outliers) **(C)** Error bar summary.

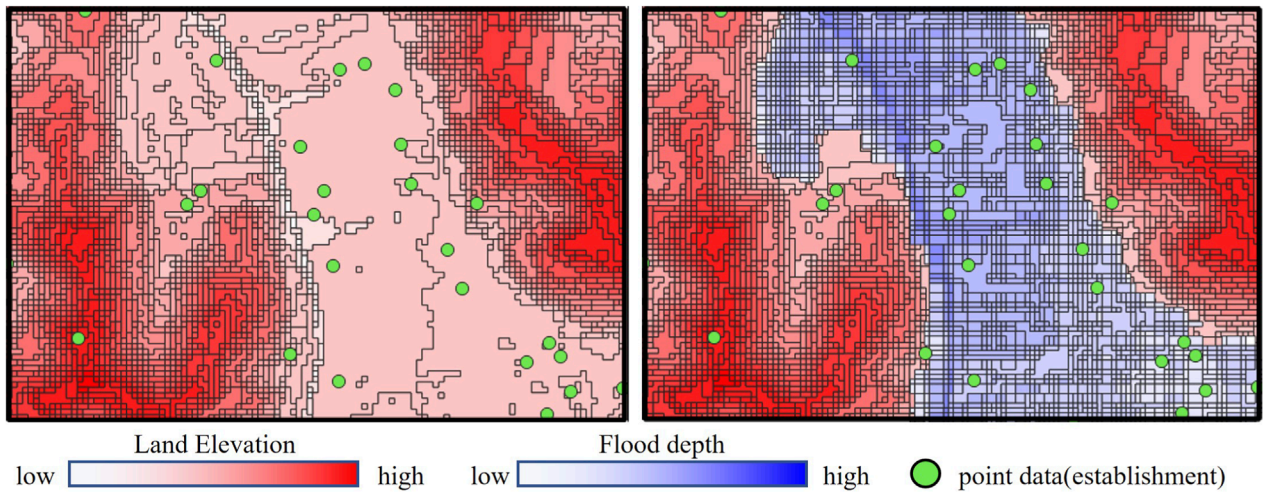
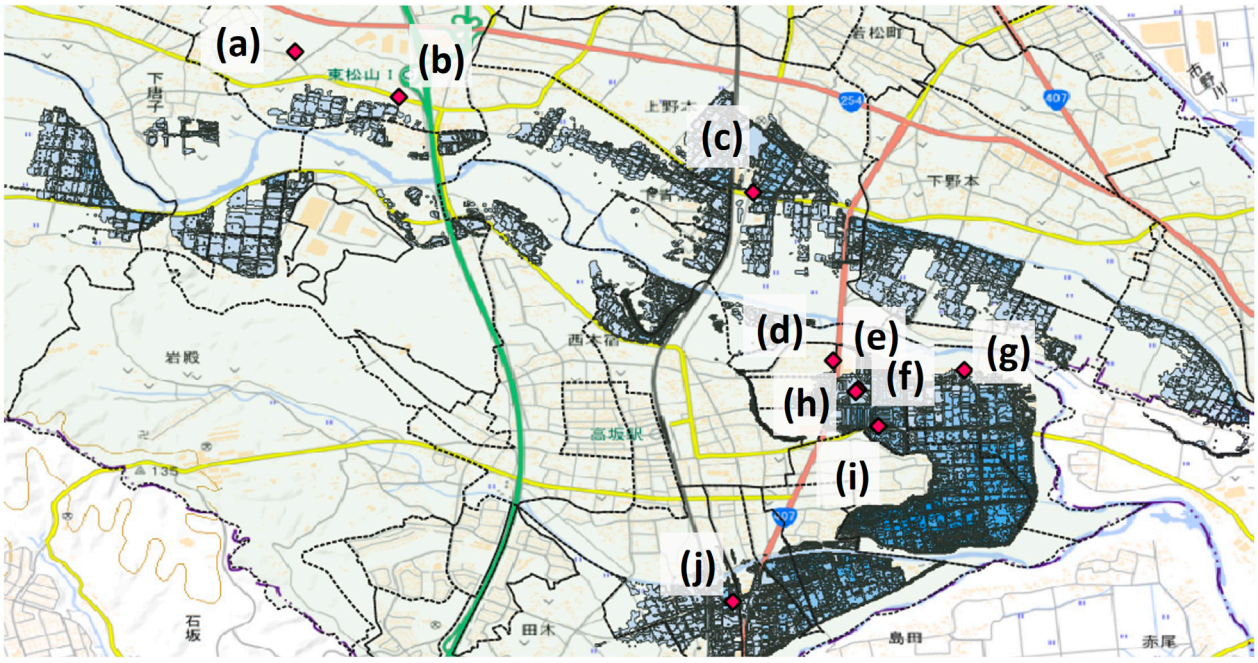


FIGURE 11
Example where point-based business damage estimates are overestimated.

consistent with the bias due to grid aggregation resolution shown in Bryant et al. (2023), where the inundation area increased by a factor of 1–2 and the number of exposed assets increased by a factor of 1–18 when the aggregation unit was set to 1–512 m. Conversely, an example of an underestimated grid could occur when only areas suitable for construction—those with significant elevation differences within the grid and low-lying, flat floodplains—become inundated. Figure 11 shows an example of a grid where flood damage was underestimated. The data depicted includes elevation data (5 m raster map) and building point data. The flooded areas (5 m raster map) are displayed in an integrated manner using QGIS. Hence, cases that could lead to serious damage may be overlooked in grid-based estimates. Therefore, it is crucial to account for the location conditions of buildings by using building point data for accurate damage estimation.

5.2 Verification of flood damage estimates: Industrial damage in Higashimatsuyama City

To verify flood damage using the building database, we compared it with the industrial damage in Higashimatsuyama City, which was extracted from a questionnaire survey (Nihei et al., 2020). Figure 12 shows the estimated flooding map (Geospatial Information Authority of Japan, 2018) of the damaged area and the locations of business establishments identified in the questionnaire survey. Figure 12A was drawn integrally using QGIS, with GSI standard maps as background maps and the coordinate system set to WGS84. The verification method involved identifying business establishments from the building database that corresponded to those in the questionnaire survey and comparing each industry with the presence or absence of flooding. The



label	Matching conditions	main	sub1	sub2
(a)	Name match	1	0	0
(b)	Name match	1	0	0
(c)	Name match	1	0	0
(d)	Address match	0	1	0
(e)	Inside the same building	1	0	0
(f)	Inside the same building	1	0	0
(g)	Name match	1	0	0
(h)	Inside the same building	0	1	0
(i)	Adjacent building	0	0	0
(j)	Adjacent building	0	0	0
accuracy		0.6	0.2	0
		0.8		

(B)

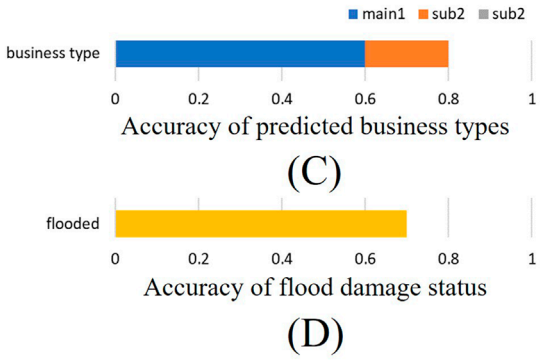


FIGURE 12
Comparison of industrial damage from questionnaire surveys and predictions **(A)** Location of affected businesses and flooded area, **(B)** Matching conditions between point data and surveys and accuracy of industry predictions, **(C)** Accuracy of predicted business types, **(D)** Accuracy of flood damage status.

identification process began by matching building names. If no matching name was found, a search was performed by address. In this search, buildings with the same address, within the same building, or adjacent buildings were identified. While adjacent buildings were not ideal for comparing industries, they were considered valuable for determining whether flooding had occurred.

An analysis of the matched data revealed an industry classification accuracy rate of 60% when only the primary industry was considered correct, and 80% when both the primary and additional industries were taken into account (Figure 12C). The remaining establishments, for which the classification was incorrect, were adjacent buildings, suggesting a high overall classification accuracy. A comparison of the flooding status revealed a 70% accuracy rate in determining whether establishments were flooded. Establishments with inconsistent answers were located in areas with large elevation differences or surrounding farmlands, suggesting that the estimated flooding map may not have fully captured the actual flooded area. A more precise understanding of the flooded area is required for accurate flood damage estimation (Katano et al., 2020). However, since the flooded area is set for various scenarios in future damage predictions, it is believed that flood damage can be accurately estimated by the establishment industry using the building database developed in this study.

6 Discussion

The following considerations can be made regarding the proposed approach based on the analysis results.

First, multimodal classification integrating an LSTM trained on business names from telephone directories with spatial information is effective for assigning business names to industry categories.

This aligns with prior research (Jiang et al., 2015), indicating that business names and surrounding environments are effective for land use and industry inference. Furthermore, combining probabilistic assignment of primary and secondary labels improved prediction stability (Figure 6). However, performance degradation is observed with variant spellings, abbreviations, rare terms, and underestimation of rare classes (Figure 7). This study constructs and evaluates only the LSTM-based multimodal classification. Yet, from the perspective of alternative model architectures, testing modern pre-trained language models like BERT could potentially further improve accuracy while keeping training costs within realistic bounds (Devlin et al., 2019). Hybrid approaches combining these models with fast, simple deep models also represent future challenges. Furthermore, considering language and institutional dependencies (e.g., Japanese terminology/domestic zoning), verifying generalizability to other languages and countries is necessary. In many high-income countries, business directories (e.g., US: NAICS-labeled commercial facility directories (Data Axle, Dun & Bradstreet, etc.); United Kingdom: Public company registers with SIC/NACE codes (Companies House, etc.) are available, potentially enabling relatively easy application. Furthermore, this workflow can be constructed from samples even without comprehensive directories. Even when establishment data is sparse, a base establishment list can be developed by combining POI platforms, OSM, and local directories. Additionally, when facility data like OSM is sparse, a minimal database can be created through targeted surveys focused on high-risk disaster areas within the analysis region.

Second, applying the classifier to building point data improves spatial detail compared to grid-based statistics and enables bottom-up aggregation from points. This allows the inferring of missing business categories from names and location information, reducing manual survey burdens (Figure 8). Accurate building location data leads to high-resolution exposure data, enabling more precise flood damage estimation (Jacquez and Rommel, 2009; Zimmerman et al., 2010; Bertsch et al., 2022) and more reliable sector-specific economic loss estimation (Yang et al., 2016; Liu et al., 2022). However, point data-specific constraints—such as coordinate errors, time lags in property records, multiple tenants within a single parcel, and dependency on industry classification rules—impact these benefits. Therefore, regular coordinate verification and sensitivity testing against alternative classification rules are required. Furthermore, this study only presents examples using the Ministry of Land (2018), Infrastructure, Transport and Tourism's 'Flood Countermeasure Economic Survey Manual (2018–2022)' for damage estimation. It has not yet been applied to advanced models estimating indirect damage or recovery processes (Darnkachatarn and Kajitani, 2025). Clarifying issues such as the relative error in damage estimates due to industry classification accuracy remains a future challenge.

Third, we quantitatively assess the impact of industrial classification errors on flood damage estimation. The correct classification rate for the multimodal industry classification is approximately 88% (Figure 6). In this workflow, by directly matching building points with phone directories for 25% of cases and estimating the remaining 75% using LSTM, the overall expected correct classification rate is $0.25 + 0.75 \times 0.88 = 0.91$. To examine the worst-case impact, we compare the manually assessed damage amounts for the smallest sector, Finance (¥1,131 thousand/person), and the largest sector, Electricity, Gas, Heat Supply, and Water Supply (¥127,803 thousand/person). The 9% error margin on the difference of ¥126,672 thousand per person between these two industries translates to an error of ±¥11,400.48 thousand per employee. Applying this result to the damage cases in City C, if applied to 64.44 employees (9% of the 716 affected employees), the error amount would be approximately ±¥734,646 million. This represents a ratio of approximately ±3.9% relative to the published damage amount of ¥18,624,453 thousand. Similarly, the maximum error range due to classification errors across the 11 regions surveyed was 0.00%–13.47%, with an average of 4.35%. The classification accuracy of the model for non-manufacturing industries, including finance, was generally good, suggesting the actual impact of classification errors may be smaller. This result is sufficiently small compared to the overestimation/underestimation range of 0%–236% calculated using a 500 m grid. Therefore, this study confirms that the business establishment database generated by this workflow can provide more appropriate damage estimates than aggregate-based data.

7 Conclusion

This study represents the first implementation of multimodal classification integrating spatial information with conventional LSTM-based text classification as a method for constructing an industrial classification-based business establishment database linked to flood damage models. Through probabilistic assignment incorporating both primary and secondary labels, the classifier achieved approximately 88% accuracy. Applying this to building point data across 11 regions and comparing it with official business statistics aggregated at the administrative district level yielded $R^2 > 0.82$ for non-manufacturing in all regions and $R^2 > 0.80$ for manufacturing in 8 regions. Bottom-up aggregation from points withstood grid statistics reconstruction; moreover, the error range in disaster estimation due to industry classification prediction accuracy proved minor compared to the variation range from grid statistics. Therefore, from an academic perspective, integrating company name-based classification into a geospatial modeling framework enhances the spatial resolution and reliability of damage estimation, helping bridge the gap between natural disaster assessment and spatial econometrics.

The limitations of this study include bias in the recorded business sectors within phone directories, potential biases in name notation, scalability beyond Japanese-speaking regions and zoning systems, and the propagation of uncertainty during the aggregation process from point to area in probabilistic classification. Future research should address extending LSTM methods, comparing alternative deep learning architectures, expanding the verification regions and disaster types, and conducting application-oriented industry evaluations. Indeed, this study only evaluates direct flood damage. To make this workflow practical, the impact of industry classification accuracy must be re-evaluated from perspectives such as its effect on indirect damage and recovery in advanced flood damage models.

In summary, this framework possesses distinct strengths enabling high-resolution disaster risk modeling while minimizing data collection burden, providing a practical foundation for recovery planning, and strengthening regional resilience.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

YKb: Writing – original draft, Writing – review and editing, Formal Analysis, Methodology, Validation. YKj: Conceptualization, Methodology, Supervision, Writing – review and editing. TU: Funding acquisition, Resources, Writing – review and editing. AY: Writing – review and editing.

Funding

The authors declare that financial support was received for the research and/or publication of this article. The authors received support from the New Energy and Industrial Technology Development Organization (NEDO) STREAM Project (P22003) for conducting this research.

Acknowledgements

This article is based on results obtained from a project, JPNP22003, subsidized by the New Energy and Industrial Technology Development Organization (NEDO). We would like to thank Editage (www.editage.jp) for English language editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbuil.2025.1631964/full#supplementary-material

References

Akiyama, Y., and Shibasaki, R. (2011). Spatio-temporal integration method for shop and office data with location information and application for urban and regional analysis. *Geogr. Inf. Syst. Assoc.* 19 (2), 57–67. doi:10.5638/thagis.19.57

Asgary, A., Anjum, M. I., and Azimi, N. (2012). Disaster recovery and business continuity after the 2010 flood in Pakistan: case of small businesses. *Int. J. Disaster Risk Reduct.* 2, 46–56. doi:10.1016/j.ijdrr.2012.08.001

Bellos, V., and Tsakiris, G. (2015). Comparing various methods of building representation for 2D flood modelling in built-up areas. *Water Resour. Manag.* 29 (2), 379–397. doi:10.1007/s11269-014-0702-3

Bertsch, R., Glenis, V., and Kilsby, C. (2022). Building level flood exposure analysis using a hydrodynamic model. *Environ. Model. & Softw.* 156, 105490. doi:10.1016/j.envsoft.2022.105490NHESS

Bhuyan, K., Westen, C. J., Wang, J., and Meena, S. (2022). Mapping and characterising buildings for flood exposure analysis using open-source data and artificial intelligence. *Nat. Hazards* 119, 805–835. doi:10.1007/s11069-022-05612-4

Bryant, S., Kreibich, H., and Merz, B. (2023). Bias in flood hazard grid aggregation. *Water Resour. Res.* 59 (9), e2023WR035100. doi:10.1029/2023WR035100

Cerri, M., Steinhausen, M., Kreibich, H., and Schröter, K. (2021). Are OpenStreetMap building data useful for flood vulnerability modelling? *Nat. Hazards Earth Syst. Sci.* 21, 643–662. doi:10.5194/nhess-21-643-2021

Chen, X., Li, H., Yu, H., Hou, E., Song, S., Shi, H., et al. (2025). Counterfactual analysis of extreme events in urban flooding scenarios. *J. Hydrology Regional Stud.* 57, 102166. doi:10.1016/j.ejrh.2024.102166

Darnkachatarn, S., and Kajitani, Y. (2025). Flood damage assessment model of industrial sectors in a megacity: derivation from business survey data in the Bangkok metropolitan region. *Int. J. Disaster Risk Reduct.* 118, 2212–4209. doi:10.1016/j.ijdrr.2025.105221

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT 2019, 4171–4186.

Dias, P., Arambepola, N. M. S. I., Weerasinghe, K., Weerasinghe, K. D. N., Wagenaar, D., Bouwer, L. M., et al. (2018). Development of damage functions for flood risk assessment in the city of Colombo (Sri Lanka). *Procedia Eng.* 212, 332–339. doi:10.1016/j.proeng.2018.01.043

Dos Santos, C., and Gatti, M. (2014). "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014. the 25th international conference on computational linguistics: technical papers*, 69–78.

Dottori, F., Figueiredo, R., Martina, M. L. V., Molinari, D., and Scorzini, A. R. (2016a). INSYDE: a synthetic, probabilistic flood damage model based on explicit cost analysis. *Nat. Hazards Earth Syst. Sci.* 16, 2577–2591. doi:10.5194/nhess-16-2577-2016

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., and Feyen, L. (2016b). Development and evaluation of a framework for global flood hazard mapping. *Adv. Water Resour.* 94, 87–102. doi:10.1016/j.advwatres.2016.05.002

Englhardt, J., de Moel, H., Huyck, C. K., de Ruiter, M. C., Aerts, J. C. J. H., and Ward, P. J. (2019). Enhancement of large-scale flood risk assessments using building-material-based vulnerability curves for an object-based approach in urban and rural areas. *Nat. Hazards Earth Syst. Sci.* 19, 1703–1722. doi:10.5194/nhess-19-1703-2019

Esparza, M., Ho, Y. H., Brody, S., and Mostafavi, A. (2025). Improving flood damage estimation by integrating property elevation data. *Int. J. Disaster Risk Reduct.* 118, 105251. doi:10.1016/j.ijdrr.2025.105251

Fuchs, S., Heiser, M., Schlögl, M., Zischg, A., Papathoma-Köhle, M., and Keiler, M. (2019). Short communication: a model to predict flood loss in mountain areas. *Environ. Model. & Softw.* 117, 176–180. doi:10.1016/j.envsoft.2019.03.026

Geospatial Information Authority of Japan (2018). Flooding estimation tiered map: heavy rain in July 2018 estimated flooding color map (aerial photo interpretation version) Takahashi River (Kurashiki city, Okayama prefecture, etc.). Available online at: https://maps.gsi.go.jp/development/ichiran.html (Accessed on January 6, 2025).

Goldblatt, R., Jones, N., and Mannix, J. (2020). Assessing OpenStreetMap completeness for management of natural disaster by means of remote sensing: a case study of three small island states (Haiti, Dominica and St. Lucia). *Remote Sens.* 12 (1), 118. doi:10.3390/rs12010118

Haque, S., Ikeuchi, K., Shrestha, B. B., Kawasaki, A., and Minamide, M. (2023). Relationship between residential house damage and flood characteristics: a case study in the Teesta River Basin, Bangladesh. *Int. J. Disaster Risk Reduct.* 96, 103901. doi:10.1016/j.ijdrr.2023.103901

Hénonin, J., Ma, H., Yang, Z. Y., Hartnack, J., Havnø, K., Gourbesville, P., et al. (2013). Citywide multi-grid urban flood modelling: the July 2012 flood in Beijing. *Urban Water J.* 12 (1), 52–66. doi:10.1080/1573062X.2013.851710

IPCC (2023). "Climate change 2023: synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental Panel on climate change," in *Core writing team*. Editors Lee, H., and Romero, J. (Geneva, Switzerland: IPCC), 35–115. doi:10.59327/IPCC/AR6-9789291691647

Ito, Y., Nakamura, S., Yoshimura, K., Watanabe, S., Hirabayashi, Y., and Kanae, S. (2019). Analysis of flood damages in the 2018 heavy rainfall with focusing on buildings location and its changing processes. *Jpn. Soc. Civ. Eng.* 75 (1), 299–307. doi:10.2208/jscejhe.75.1_299

Jacquez, G. M., and Rommel, R. (2009). Local indicators of geocoding accuracy (LIGA): theory and application. *Int. J. Health Geogr.* 8, 60. doi:10.1186/1476-072X-8-60

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., and Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst.* 53, 36–46. doi:10.1016/j.compenvurbsys.2014.12.001ResearchGate

Katano, Y., Akamatsu, I., Tamara, S., and Tanaka, T. (2020). Study on the characterisitcs of building damages caused by sediment disaster and flood disaster of the heavy rain event of July 2018 - analysis using disaster victim certificate data of Mihara city in Hiroshima prefecture. *J. City Plan. Inst. Jpn.* 55 (3), 851–857. doi:10.11361/journalcpij.55.851

Kim, Y. (2014). "Convolutional neural networks for sentence classification," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1746–1751. doi:10.3115/v1/D14-1181

Kuroda, N., Kajitani, Y., and Tatano, H. (2020). Estimating fragility curves for asset damage in business sector caused by a flood disaster: a case of the heavy rain event of July 2018. *Jpn. Soc. Civ. Eng.* 76 (1), 70–80. doi:10.2208/jscejhe.76.1_70

Lee, J. Y., and Dernoncourt, F. (2016). "Sequential short-text classification with recurrent and convolutional neural networks," in Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: human language technologies, 515–520. doi:10.18653/v1/N16-1062

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., et al. (2022). A survey on text classification: from traditional to deep learning. *ACM Trans. Intelligent Syst. Technol.* 13 (2), 1–41. Article 31. doi:10.1145/3495162

Liu, H., Tatano, H., Kajitani, Y., and Yang, Y. (2022). Analysis of the influencing factors on industrial resilience to flood disasters using a semi-markov recovery model: a case study of the heavy rain event of July 2018 in Japan. *Int. J. Disaster Risk Reduct.* 82, 103384. doi:10.1016/j.ijdrr.2022.103384

Ma, J., Blessing, R., Brody, S. D., and Mostafavi, A. (2024). Non-locality and spillover effects of residential flood damage on community recovery: insights from high-resolution flood claim and mobility data. *Sustain. Cities Soc.* 117, 105947. doi:10.1016/j.scs.2024.105947

Maki, S., Ohnishi, S., Fujii, M., and Goto, N. (2023). "Development of the classification model for steam demand factories and estimation of spatial steam demand by text analysis using company names in telephone book data," in *Proceedings of the 42nd Annual Meeting of the Japan Society of Energy and Resources*, 45, 145–146.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A. (2010). Review article "Assessment of economic flood damage". *Nat. Hazards Earth Syst. Sci.* 10, 1697–1724. doi:10.5194/nhess-10-1697-2010

Ministry of Internal Affairs and Communication (2023). Japan standard industrial classification. Available online at: https://www.soumu.go.jp/toukei_toukatsu/index/seido/sangyo/index.htm?entryAnkenIds=49565 (Accessed on October 10, 2024).

Ministry of Internal Affairs and Communications (2014). Economic census for business frame. Available online at: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200552&metadata=1&data=1 (Accessed on October 10, 2024).

Ministry of Internal Affairs and Communications (2015). Population census. Available online at: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200521&metadata=1&data=1 (Accessed on September 24, 2024).

Ministry of Internal Affairs and Communications (2016). Economic census for business activity. Available online at: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200553&metadata=1&data=1 (Accessed on October 10, 2024).

Ministry of Internal Affairs and Communications (2018). Housing and land survey. Available online at: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00200522&metadata=1&data=1 (Accessed on November 15, 2024).

Ministry of Land, Infrastructure, Transport and Tourism (2018-2022a). Flood damage statistics survey. Available online at: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00600590&result_page=1 (Accessed on June 28, 2024).

Ministry of Land, Infrastructure, Transport and Tourism (2018-2022b). "Flood control economic survey manual (draft), (National Water Management Survey Various asset valuation unit prices and deflators). Available online at: https://www.mlit.go.jp/river/basic_info/seisaku_hyouka/gaiyou/hyouka/hyouka.html (Accessed on June 28, 2024).

Ministry of Land, Infrastructure, Transport and Tourism (2019). Land use zoning. Available online at: https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A29-v2_1.html (Accessed on June 28, 2024).

Neffke, F., Henning, M., and Boschma, R. (2011). How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Econ. Geogr.* 87 (3), 237–265. doi:10.1111/j.1944-8287.2011.01121.x

Nihei, Y., Nakaegawa, T., Nakakita, H., Takemi, T., Yamada, T., Misumi, R., et al. (2020). "Comprehensive research on wide-area disasters caused by Typhoon No. 19 and Typhoon No. 21 in 2019," in *Natural Disaster Science Symposium Proceedings*, 57, 9–22.

Nippon Software Service (2022). Electronic telephone directory data. Available online at: https://www.nipponsoft.co.jp/solution/denshi30/.

Ohara, M., Nagumo, N., and Shinya, T. (2022). Analysis of recovery of enterprises after torrential rainfall disaster in July, 2018. *Jpn. Soc. Civ. Eng.* 78 (2), 37–42. doi:10.2208/jscejhe.78.2_I_37

Oubennaceur, K., Chokmani, K., Nastev, M., Lhissou, R., and El Alem, A. (2019). Flood risk mapping for direct damage to residential buildings in Quebec, Canada. *Int. J. Disaster Risk Reduct.* 33, 44–54. doi:10.1016/j.ijdrr.2018.09.007

Paulik, R., Wild, A., Zorn, C., and Wotherspoon, L. (2023). Bias in flood hazard grid aggregation. *Water Resour. Res.* 59 (9), e2023WR035100.

Polverino, S., Nia, H. A., and Rahbarianyazd, R. (2024). Design advances in urban impervious surface extraction: leveraging K-distributions, GLCM, Rayleigh and Nakagami with VHR SAR technology. *New Des. Ideas* 8 (Special Issue), 1–23. doi:10.62476/ndisi.01

Ran, J., and Nedovic-Budic, Z. (2016). Integrating spatial planning and flood risk management: a new conceptual framework for the spatially integrated policy infrastructure. *Comput. Environ. Urban Syst.* 57, 68–79. doi:10.1016/j.compenvurbsys.2016.01.008

Scorzini, A. R., and Frank, E. (2017). Flood damage curves: new insights from the 2010 flood in Veneto, Italy. *J. Flood Risk Manag.* 10 (3), 381–392. doi:10.1111/jfr3.12163

Teo, C. J., Poinapen, J., Hofman, J. A. M. H., and Wintgens, T. (2025). Assessing water dependencies and risks in Dutch industries: distribution, consumption and future challenges. *Water Resour. Industry* 33, 100279. doi:10.1016/j.wri.2025.100279

Tojo, T., and Oyama, Y. (2022). A deep learning model for building estimation based on building names – application to a micro land use analysis. *J. City Plan. Inst. Jpn.* 57 (3), 1025–1032. doi:10.11361/journalcpij.57.1025

Tokyo Shoko Research, Ltd. (2024). TSR corporate information file. Available online at: https://www.tsr-net.co.jp/service/detail/file-corporate.html.

Tu, J., Wen, J., Yang, L. E., Reimuth, A., Young, S. S., Zhang, M., et al. (2023). Assessment of building damage and risk under extreme flood scenarios in Shanghai. *Nat. Hazards Earth Syst. Sci.* 23, 3247–3260. doi:10.5194/nhess-23-3247-2023

Ullah, T., Lautenbach, S., Herfort, B., Reinmuth, M., and Schorlemmer, D. (2023). Assessing completeness of OpenStreetMap building footprints using MapSwipe. *ISPRS Int. J. Geo-Information* 12 (4), 143. doi:10.3390/ijgi12040143

Yang, L., Kajitani, Y., Tatano, H., and Jiang, X. (2016). A methodology for estimating business interruption loss caused by flood disasters: insights from business surveys after Tokai heavy rain in Japan. *Nat. Hazards* 84, 411–430. doi:10.1007/s11069-016-2534-3

Zabret, K., Hozjan, U., Kryžanowsky, A., Brilly, M., and Vidmar, A. (2016). Development of model for the estimation of direct flood damage including the movable property. *J. Flood Risk Manag.* 11 (S1), S527–S540. doi:10.1111/jfr3.12255

Zeng, Z., Guan, D., Steenge, A. E., Xia, Y., and Mendoza-Tinoco, D. (2019). Flood footprint assessment: a new approach for flood-induced indirect economic impact measurement and post-flood recovery. *J. Hydrology* 579, 124204. doi:10.1016/j.jhydrol.2019.124204

Zenrin Co., Ltd. (2018). Building point data:kurashiki city (2018), Ozu city (2018), Higashimatsuyama city (2018), Mobara city (2019), Kakuta city (2019), Koriyama city (2020), Tikuma city (2020), Omuta city (2020), Hitoyoshi city (2020), Musashio city (2021), Murakami city. Available online at: https://www.zenrin.co.jp/product/category/gis/contents/building-point/index.html.

Zhu, J., and Cao, Y. (2024). "A study of fusion strategies for self-attention enhanced convolutional neural networks in short text classification," in 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE) (Guangzhou, China), 587–592. doi:10.1109/CISCE62493.2024.10653453

Zimmerman, D. L., Fang, X., Mazumdar, S., and Rushton, G. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *Int. J. Health Geogr.* 9, 10. doi:10.1186/1476-072X-9-10