scientific data

OPEN

DATA DESCRIPTOR

MSLU-100K: A Large Multi-Source Dataset for Land Use Analysis in Major Chinese Cities

Yao Yao[1,2,3,4,5] ✉, Yueheng Ma[1], Ronghui Gao[1], Xiaoqin Yan[6,7] & Qingfeng Guan[1,2]

High-quality land use datasets are essential for advancing research in land use classification and recognition. However, the complexity and spatial heterogeneity of land use create challenges in dataset construction. To address these issues, we present MSLU-100K, a multi-source land use dataset encompassing over 100,000 irregular parcel samples from 81 Chinese cities. Constructed using a human-computer collaboration framework, this dataset integrates remote sensing and POI (Point of Interest) data, categorizing parcels into 7 primary and 28 secondary land use types. A novel multi-level classification approach combines manual labeling and deep learning, ensuring high data quality across six quality levels. Over 57% of the dataset comprises high-quality samples (Levels 4 and 5), which significantly enhance classification performance. The dataset provides a robust resource for land use recognition, urban planning, and spatial research.

Background & Summary

As a crucial foundation for urban planning and sustainable development, accurate land use classification effectively reflects regional socio-economic contributions[1]. It provides essential support for analyzing the impacts of land use changes on the ecological environment[2]. With the development of deep learning and geospatial big data, the capability to mine and integrate natural physical and socio-economic attributes from multi-source spatiotemporal data makes precise land use identification feasible[3–5].

In constructing machine learning models, significant time is often spent preparing training data, as data quality is critical to a model's overall performance[6]. Consequently, high-quality land use datasets are essential for building high-performance land use classification models. Land use classification involves three main stages: data sampling, data labeling, and model training. With the accelerated pace of global urbanization, urban areas are constantly expanding, resulting in increasingly complex land use types[7–9]. The spatial heterogeneity of land use data, along with the easily confusable nature of land use types make data sampling, data labeling, and model training difficult.

During the data sampling stage, the modifiable areal unit problem (MAUP)[10,11] emerges due to multi-scale parcels, necessitating that model training incorporates multi-scale data to ensure model robustness. Furthermore, the spatial heterogeneity of land use[12] complicates the direct transfer of knowledge across regions, making it essential to account for spatial distribution characteristics during sampling. In the data labeling phase, effective visual interpretation of land use demands geographic expertise. Additionally, the numerous land use categories, with many prone to confusion[13], create challenges in constructing land use datasets through visual interpretation or artificial intelligence-based automatic labeling, thus impacting both the efficiency and quality of dataset development. During model training, the significant imbalance among land use categories[14,15] presents challenges for training machine learning models[16]. Moreover, mixed land use categories are very common in

[1]UrbanComp Lab, School of Geography and Information Engineering, China University of Geosciences, Wuhan, 430078, Hubei province, China. [2]National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, 430078, Hubei province, China. [3]LocationMind Institution, LocationMind Inc., Chiyoda, Tokyo, Japan. [4]Hitotsubashi Institute for Advanced Study, Hitotsubashi University, Kunitachi, Tokyo, Japan. [5]Faculty of Engineering, Reitaku University, Kashiwa, Chiba, Japan. [6]Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing, 100871, China. [7]Beijing Key Lab of Spatial Information Integration & Its Applications, Peking University, Beijing, China. ✉e-mail: yaoy@cug.edu.cn; yao.yao@urbancomp.net

Dataset Name	Year	Number of Land Use Categories	Region
PatternNet	2018	38	America
NWPU-RESISC45	2017	45	Global
EuroSAT	2019	34	Europe
ILU- CUG	2022	20	Four first-tier and second-tier cities in China

Table 1. Existing land use datasets.

modern cities[17], but they cannot be clear labeled and cannot be directly used as training datasets. Consequently, the quality of land use data presents severe challenges to the accuracy of classification models.

Massive public datasets such as images, text and audio have greatly facilitated machine learning by training and optimizing models. Specifically, land use datasets contain a wealth of geographical information such as land types, terrain, and vegetation, which are invaluable for studying land use patterns and trends or for training land use classification models. Numerous scholars have released open datasets for land use classification, as illustrated in Table 1. Examples include PatternNet[18], NWPU-RESISC45[19], EuroSAT[20], and ILU-CUG[21]. Open datasets greatly aid researchers by providing access to comprehensive experimental data. These datasets vary in data volume, number of labels, and regional focus, which can lead to inconsistencies in quality. Such variations can significantly affect the performance of land use classification models, making accurate assessment of dataset quality a critical issue for researchers[22].

Existing studies have shown that high quality training data can enhance deep learning performance[23–25]. Thus, a systematic evaluation of dataset quality is essential for building high-performance machine learning systems. To ensure dataset effectiveness and reliability across different research tasks, researchers employ various methods to assess dataset quality. Subjective evaluation methods assess data quality through volunteers based on their own subjective perceptions. For instance, Mohammadi *et al.*[26] discussed various subjective data quality evaluation methods, but their existing limitations also prompted researchers to continuously explore more efficient and economical alternative solutions to meet the growing application demands. NEHMÉ *et al.*[27] conducted a large-scale crowdsourcing experiment, engaging over 4,500 participants to rate distorted images. However, the results obtained from subjective methods can be influenced by the participants' subjective experiences and emotions, as well as higher implementation costs.

In recent years, with the development of deep learning technology[28], data quality evaluation has become more efficient. Chen *et al.*[29] introduced a remote sensing image quality assessment framework that predicts the quality score of each image by extracting image features and employing a 3D CNN model. Kang *et al.*[30] modified the standard CNN architecture to directly learn and predict image quality from raw image pixels. Nevertheless, the evaluation criteria for these objective methods remain unclear, providing only binary labels of "good" and "bad", and lacking a detailed delineation of quality dimensions. Consequently, clear quality grade divisions for datasets are challenging. It is necessary to integrate manual judgment with deep learning methods to achieve a more accurate and comprehensive assessment of dataset quality.

To tackle the issues mentioned above, this study used a human-machine collaborative labeling method to construct a multi-source sample dataset of land use in major cities in China (MSLU-100K). Leveraging this dataset, we propose a multi-level model classification method based on manual filtering and a grading method based on model soft classification probability to evaluate and classify the quality grades of the land use dataset. By comparing model accuracy trained with datasets of varying quality grades, we validate the dataset's usability.

As one of the largest datasets in the field of land use for major Chinese cities, the MSLU-100K dataset focuses on land use samples of irregular parcels, integrating multi-source data such as remote sensing images and POI, and exhibits strong spatial heterogeneity. This comprehensive and diverse dataset, attentive to specific parcel characteristics, offers a robust foundation for enhancing the accuracy and stability of land use classification. Moreover, the quality evaluation method effectively addresses previous shortcomings in dataset quality assessment standards, especially regarding the comprehensiveness and accuracy of evaluation.

Methods

The workflow of constructing our land use dataset is shown in Fig. 1. Initially, we gather the data necessary for constructing the land use dataset. Subsequently, we employ the DCAI-CLUD implementation method, which includes sample filtering based on parcel location and size, as well as a "human-computer collaboration" approach for dataset construction[31]. Finally, we assess and validate the dataset quality through the integration of manual judgment and deep learning methods.

Data collection and processing.

This section summarizes the data products used for constructing the land use dataset and quality assessment, along with the corresponding preprocessing steps. Table 2 provides details on the type, source, and function of all datasets.

Unlabeled parcel boundary data. The unlabeled parcel boundary data, also called the Area of Interest (AOI) data, The generation of the parcel data used two datasets: the administrative division data of each city downloaded through the Alibaba Cloud Data Visualization Platform and the road network data downloaded from OpenStreetMap (OSM). For road network data at each level, buffer zones were established based on roadbed width. The QGIS vector clipping tool was employed to clip road buffer zones from the administrative division

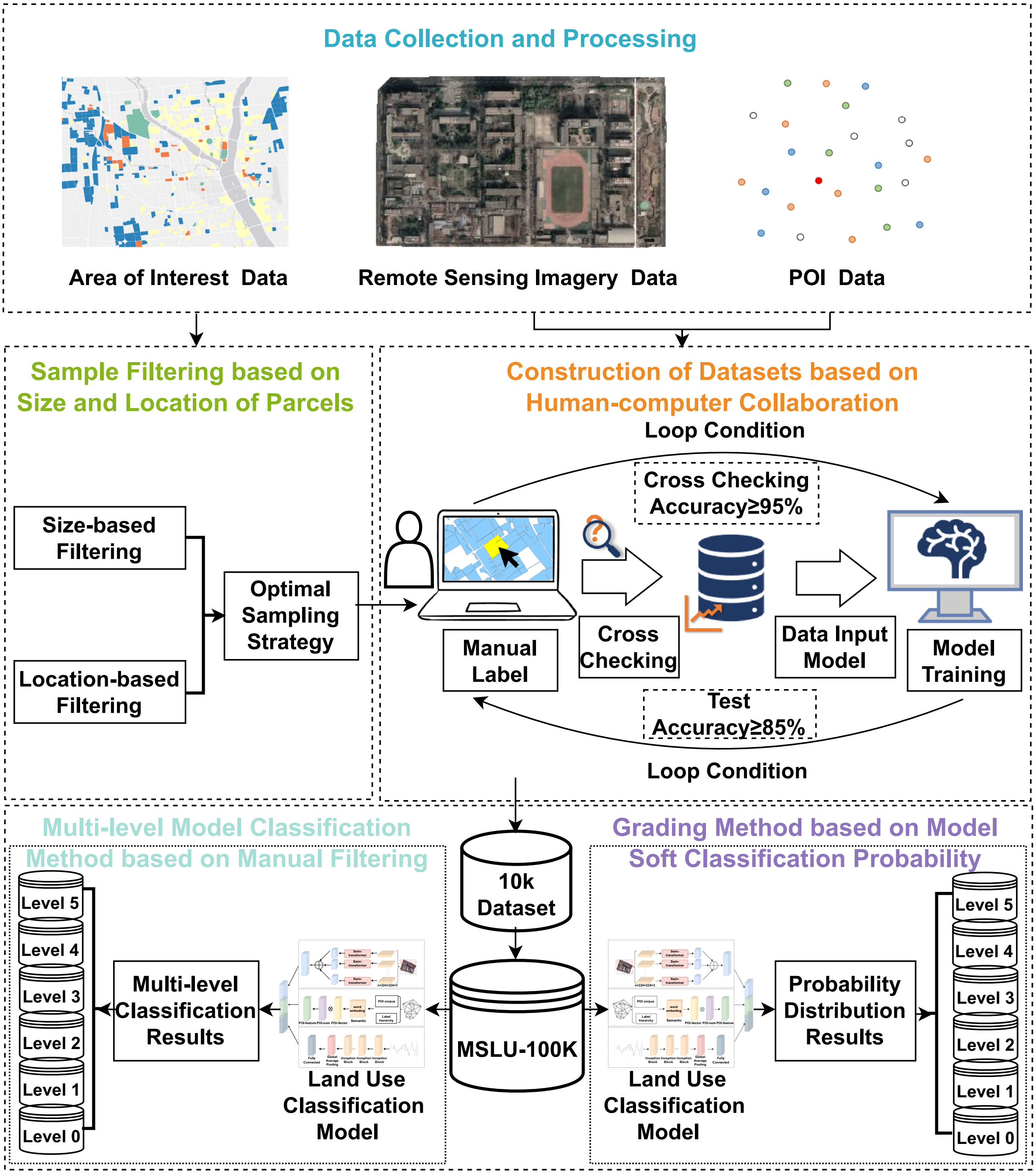
Data Collection and Processing
Area of Interest Data
Remote Sensing Imagery Data
POI Data
Sample Filtering based on Size and Location of Parcels
Size-based Filtering
Location-based Filtering
Optimal Sampling Strategy
Construction of Datasets based on Human-computer Collaboration
Loop Condition
Cross Checking Accuracy≥95%
Manual Label
Cross Checking
Data Input Model
Model Training
Test Accuracy≥85%
Loop Condition
Multi-level Model Classification Method based on Manual Filtering
10k Dataset
MSLU-100K
Grading Method based on Model Soft Classification Probability
Level 5
Level 4
Level 3
Level 2
Level 1
Level 0
Multi-level Classification Results
Land Use Classification Model
Land Use Classification Model
Probability Distribution Results

Fig. 1 Overall Technical Roadmap.

polygon vector data. Subsequently, a positive and negative 5-meter buffer was applied to the clipped data results, culminating in a total of 409,800 parcels across the country.

Multi-source spatiotemporal data. We developed a multi-source land use recognition deep learning model by integrating remote sensing images with point of interest data. This trained model was then employed for automatic labeling during the dataset construction phase. Moreover, we utilized an API provided by Tencent to acquire user density data from May 7, 2019. The incorporation of Tencent's user density data aimed to enhance the model's effectiveness in recognizing residential, commercial, and public land use areas with significant variations in human traffic[32].

Sample filtering based on size and location. The land use data employed in this study comprises irregular parcels generated by road networks. Overlapping road network data results in some parcels being too small

Data Name	Years	Data Type	Data Description
Chinese Administrative Divisions	2022	Vector	Administrative Division Data of Cities in China (https://datav.aliyun.com)
OSM Road Network	2022	Vector	OpenStreetMap (OSM) Road Network (https://www.openstreetmap.org)
Unlabeled parcel boundary data	2023	Vector	Parcel boundaries are generated using administrative divisions and road network data by employing a road buffer clipping and morphological optimization method.
High-Resolution Visible Light Remote Sensing Images	2019–2022	Raster	The data source comes from Google Earth Engine(Google Earth Engine, GEE) created by merging and stitching multi-source data. The time frame is from 2019 to 2022, with a spatial resolution of 3 meters.
Gaode POI	2018	Vector	Reflects socio-economic attributes and the functional structure of the city, it includes 23 primary categories and 261 secondary categories. (https://developer.amap.com/api)
Tencent User Density Data	2019	Raster	The Tencent user density data from May 7, 2019, has a temporal resolution of 1 hour and a spatial resolution of approximately 1100 meters.

Table 2. Data information used in the dataset experiment.

and others too large due to incomplete road network data, rendering them as noise in model training. Moreover, according to von Thünen's "Land Rent Theory", the size of a parcel impacts its land use category by influencing its expected profit[33]. Consequently, we identified an optimal size range to filter out noise data, mitigate category imbalance in the dataset, and reduce the proportion of parcels with mixed land use categories. Land use in urban areas is more complex and diversified than in non-urban areas, yet the complexity of land use mix escalates with urban development[17]. To acquire more varied land use data in urban areas while maximizing the quality of the dataset, we employed a data screening method grounded in parcel attributes[31]. By adjusting the dispersion coefficient, we can control the likelihood of selecting parcels at varying distances from the city center, thereby ensuring that the dataset achieves an optimal geographical distribution. This adjustment balances the impacts of random and distance factors during sampling to optimize the spatial distribution characteristics of parcels. Ultimately, we identified the optimal range for area selection (38931.315 m^2 ~ 676818.47 m^2) to eliminate parcels that are excessively small or large, thereby reducing class imbalance within the dataset and lowering the proportion of heterogeneous land-use samples. Regarding location selection, we discovered that setting the weight ratio of the distance factor to the city center and the random distribution factor to 0.7:0.3 yields the best filtering results. Under these conditions, while ensuring balanced coverage of both the urban core and peripheral areas, the selected samples not only efficiently exclude noise data that are either too small or large but also significantly reduce the proportion of mixed land-use samples. Consequently, the class balance within the dataset and the coverage rate across urban areas both reached optimal levels.

Irregular parcels land use classification model.

This study develops a multi-source deep learning model for land use identification that deeply integrates remote sensing imagery, POI, and temporal population data. The framework of the proposed model is shown in Fig. 2. The model captures external physical and internal socioeconomic characteristics to assess and validate dataset quality. Initially, Poisson disk sampling is employed to convert irregular parcels of varying shapes and sizes into multiple fixed-size parcels, which are saved as pickle files, a common format for data storage in Python. Each parcel is ultimately encoded into its corresponding pickle file. The model leverages a remote sensing image feature module based on the Swin-Transformer[34], in conjunction with fixed-size images generated via Poisson disk sampling. It also introduces a remote sensing image feature fusion layer following the methodology by Srivastava *et al.*[35]. In this study, feature extraction is accomplished using Swin-Transformer, followed by element-wise averaging for feature fusion. The POI feature extraction component, grounded in the Semantic method[36], generates training corpora for the POI network through random walks. It then utilizes the skip-gram algorithm from NLP to learn low-dimensional semantic embeddings of POI categories. This constructs a mapping matrix between POI categories and their representation vectors to extract socially-aware data features. The temporal feature extraction module, based on InceptionTime[37], comprises multiple stacked Inception Block modules optimized through residual connections for semantic information extraction from temporal data. Finally, the model assigns specific weights to each information type based on parcel characteristics, integrating POI, remote sensing imagery, and temporal feature vectors through weighted fusion. These are then processed through a fully connected layer and a SoftMax layer for classification, producing the final output.

Construction of datasets based on human-machine collaboration.

Leveraging the predictive capabilities of machine models to assist manual labeling serves as an effective means to enhance labeling efficiency[38]. Wu *et al.*[31] proposed a "human-machine collaboration" methodology based on model-assisted pre-labeling for dataset construction. The land use classification dataset is progressively constructed and refined through an iterative process that integrates manual labeling with model predictions. During the dataset construction process, we utilized only the remote sensing image feature extraction module and the POI data feature extraction module within the model framework to perform feature fusion, enhancing the effectiveness of data quality. This process continues until the desired data volume and model accuracy criteria are satisfied.

We initially implemented random sampling labeling throughout the entire Yangtze River Delta region and executed comprehensive labeling in certain key areas to ensure both the comprehensiveness and high precision of the data. Utilizing this detailed dataset, we performed analyses of samples based on location and area,

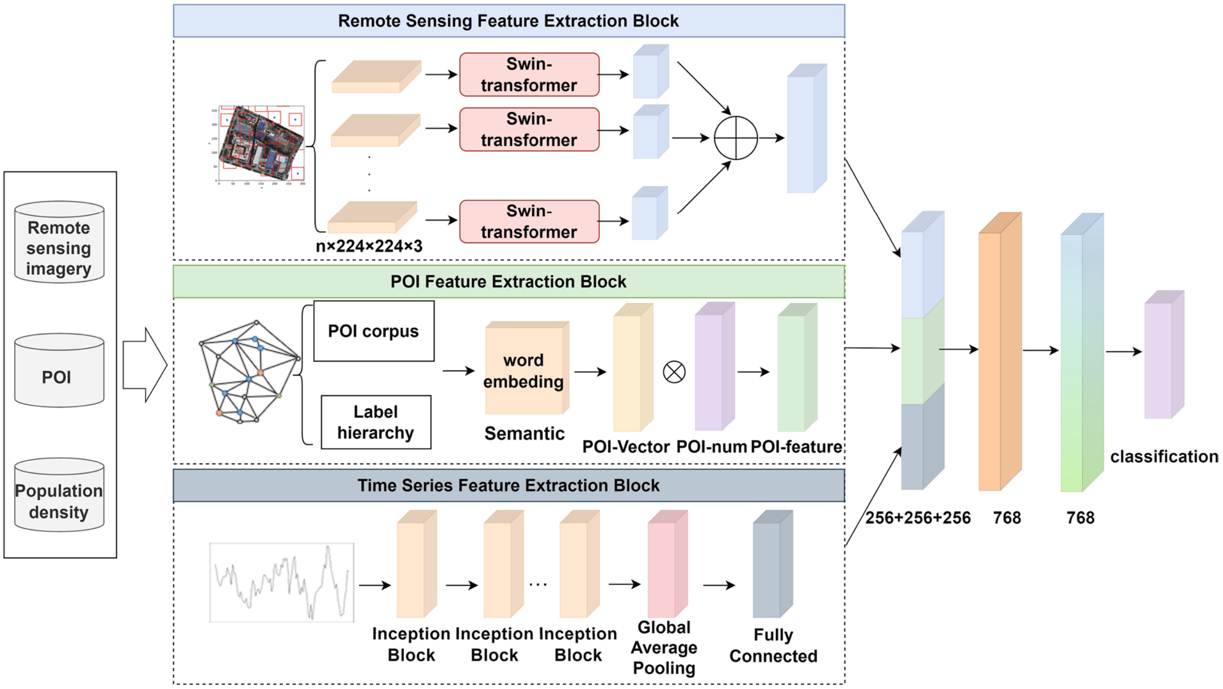

Fig. 2 Framework of the proposed model.

ultimately identifying the optimal sampling strategy[31]. Subsequently, labeling was extended nationwide, adopting the optimal sampling strategy ascertained from the Yangtze River Delta area. This approach focuses more on sampling in urban regions, ensuring a detailed representation of land use characteristics in highly urbanized areas. Moreover, during the labeling process, a "grouping learning strategy" is employed, where annotators are grouped according to categories and only annotate data predicted by the model as belonging to the same category. This approach reduces learning costs, enhances labeling efficiency, and improves understanding of the definitions of land use categories. Simultaneously, all labeling results are subject to a cross-check mechanism to ensure data quality. Annotators are required to select 25% of the data for mutual verification. If the labeling accuracy is below 90%, revisions must be made until the standard is met. In this study, the final acceptance rate for cross-checking is set at 95% to ensure the quality of the dataset.

This study constructed the datasets necessary for multi-level models through manual labeling. Irregular parcels land use classification model are employed to enhance labeling efficiency, decrease the workload of manual labeling, and assist in identifying samples that are difficult to determine. The research team trains an initial land-use classification model based on the initial manual labeling data. The prediction results are then stored in the fields of the dataset awaiting labeling, providing reference for the subsequent round of manual labeling. Labelers first verify the model's predictions, eliminating the need to start labeling from scratch and thereby accelerating the data construction process. Following each round of manual labeling, the new labeling data is used to retrain the land-use classification model, continuously optimizing the model's performance. After each iteration of model optimization, the updated model is utilized for pre-labeling the next round of data, thus forming a "human-machine collaborative optimization" closed loop. Initially, core team members meticulously labeled data to create a small-scale, high-quality dataset containing 1,000 entries. We ensured complete accuracy in labeling by integrating remote sensing images and POI data, establishing a 100% accurate standard benchmark for subsequent labeling efforts. Building on this, the research team enlisted external experts to undertake large-scale labeling in multiple rounds. The team utilized the dataset of 1,000 entries to train a deep learning model for predictions and sample selection. This approach progressively enhanced the land use classification model's accuracy, resulting in a high-quality dataset of 10,000 entries with a labeling accuracy of 95%.

To further expand the dataset, 56 labelers were recruited for additional labeling tasks. To maintain high-quality labeling across the large-scale dataset, labelers were mandated to sample 25% of their labeled data for cross-verification with fellow labelers. If a labeling task's accuracy fell below 90%, labelers were required to revise their work until surpassing the 90% threshold. Through repeated iterations of this process, the land use classification dataset was gradually refined to meet target data volume and accuracy standards. Eventually, labeled results were achieved for approximately 100,000 entries. These entries comprise 40,682 residential entries (Res), 6,286 public service land entries (Pub), 6,684 commercial entries (Com), 24,498 industrial entries (Ind), and 21,411 agricultural natural land parcels (Agr). The final dataset was named MSLU-100K (China Multi-Source Land Use Dataset).

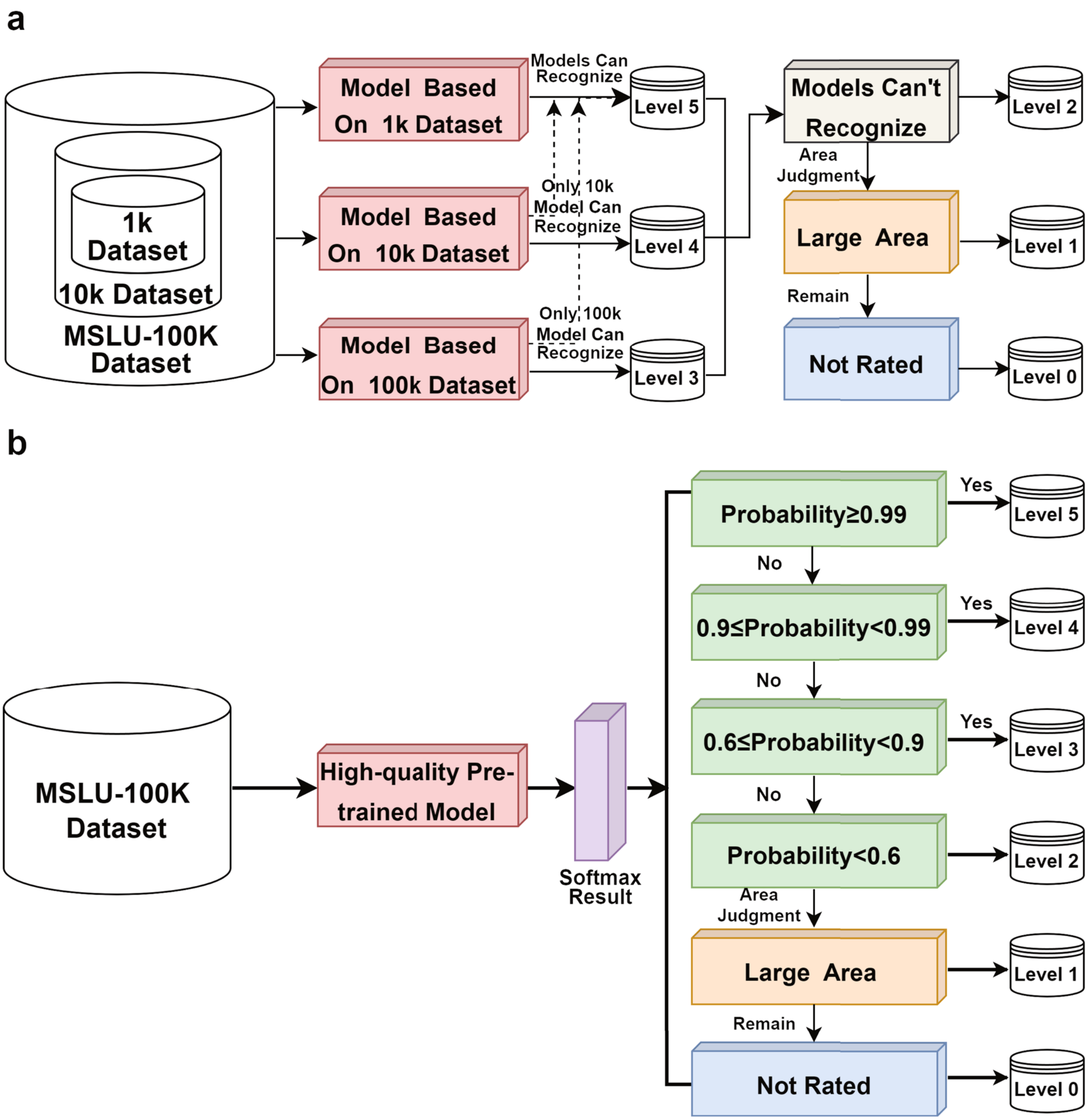

Fig. 3 Methods for dataset quality assessment (**a**) Multi-level model classification method based on manual screening, (**b**) Grading method based on model soft classification probability.

Dataset quality assessment. This study proposes two methods for evaluating the quality of land use datasets, aimed at comprehensively assessing dataset quality to clearly define dataset grading. During the data quality assessment process, we utilized all modules within the model framework, including the remote sensing image feature extraction module, the POI data feature extraction module, and the time-series population feature extraction module for feature fusion to enhance the model's performance, thereby enabling better classification of data quality. Among these, the multi-level model classification method based on manual filtering accurately grades quality by precisely capturing features through the model. Meanwhile, the grading method based on model soft classification probability uses probability distribution to quantitatively assess the purity of parcel characteristics. Both methods categorize MSLU-100K into six levels, ranging from 0 to 5.

Figure 3(a) illustrates the multi-level model classification method based on manual filtering. Models trained on high-quality data capture target features more accurately, enhancing the precision of machine learning classification tasks[6]. We constructed datasets of scales 1k, 10k, and 100k based on manual labeling rules and subsequently trained three land use classification models with varying complexities. Starting with the most rigorously filtered data (1k), we gradually extended the datasets to ensure that the high-quality data were used to train high-precision models.

Initially, we utilized 1,000 rigorously selected high-quality data points to train the 1k level model. Since these samples have complete Points of Interest (POI) information, clear boundaries, and no mixed land use, the model requires only a few features to classify accurately. Samples that can be accurately classified by this model are deemed to be of the highest quality, marked as level 5. For samples that the 1k level model cannot classify, we used the 10k level model, trained with 10,000 extended high-quality data points. If this model can classify these data accurately, it indicates that although the complexity is slightly higher, the data are still high quality, rated as level 4. When the 10k level model still fails to classify the data, we employ the 100k level model, trained on the complete MSLU-100K dataset, for further identification. At this stage, data quality is further reduced, and samples are rated as level 3. If the 100k level model also cannot classify them correctly, or manual inspection determines the sample quality is poor, the samples are categorized into low-quality levels, including level 2 (significant mixed land use, high classification difficulty), level 1 (unable to be judged manually), and level 0 (lacking POI or remote sensing information, unusable for classification).

The rationale behind this grading method lies in the fact that the 1k level model's training data are the most stringent, containing only clear, single land parcels; thus, samples correctly classified by it are of the highest quality. In contrast, the 10k and 100k models use training data of increasing complexity, which results in increased classification uncertainty, and therefore the quality of samples corresponding to their classification results is relatively lower. Ultimately, through model classification capability evaluation and manual screening, we achieve efficient dataset grading, ensuring the reliability of the data quality, while filtering out low-quality samples to enhance the overall value of data utilization.

Figure 3(b) illustrates the grading method based on model soft classification probability. In this study, the pre-trained high-precision 100k model is utilized to derive the soft classification results for the MSLU-100K dataset. Softmax[39] is employed to calculate the probability distribution for each category, ensuring the sum of probabilities equals 1. Based on these probability distribution results, data is graded with a higher probability suggesting that the model perceives the parcel sample as having no other functional characteristics. This indicates that the parcel possesses clear and distinct functions, reflecting superior data quality, while quality decreases from level 5 to level 0. A probability of 0.99 or higher indicates over a 99% chance of being identified as a specific category, which is rated as level 5. Probabilities between 0.9 and 0.99 are rated as level 4; probabilities between 0.6 and 0.9 are rated as level 3; probabilities below 0.6 are rated as level 2; parcels too large for human recognition are rated as level 1; and samples that lack label and POI information, reflecting the lowest quality, are rated as level 0.

Data Records

This section introduces the dataset structure and format, which comprises two folders, a Python program, and a CSV file. The Classification folder contains metadata files in XML format for samples, including information such as sample category, path, and image size. The ImageSets folder houses remote sensing images categorized by land use types, divided into Agr, Res, Com, Pub, and Ind. The DatasetGenerate.py file provides sample code for generating dataset tables from XML files. Executing this script results in the creation of MSLU-100K.csv, the dataset table. This table includes details on category, file name, storage path, image width, image height, geographic information, primary category name, and secondary category name for all entries. The dataset[40] is publicly available for free on Open Science Framework (https://doi.org/10.17605/OSF.IO/YAENR).

Technical Validation

The validation of this study comprises three sections: (1) statistical results of dataset quality assessment; (2) evaluation of model performance; (3) analysis of land use mapping.

Statistical results of dataset quality assessment. The distribution of quality level results derived from the two evaluation methods across different categories are presented in Tables 3 and 4. Overall, across the two methods, level 4 and 5 data comprise approximately 57.1% of the dataset, excluding level 0 data. Furthermore, level 5 data accounts for more than 40% of the data set and level 4 data accounts for more than 10%, both of which represent a substantial proportion of the total dataset. These findings demonstrate the high reliability and credibility of the dataset constructed in this study regarding data quality. Further analysis reveals that these high-quality data are predominantly found in categories such as residential, agricultural, and industrial sectors. This suggests that the data in these categories are of relatively better quality compared to others, likely due to their distinct characteristics and higher consistency. The multi-level model classification method based on manual filtering considers the model's recognition capability as the evaluation criterion, resulting in a relatively high proportion of unidentified level 2 data. In contrast, the grading method based on the model soft classification probability employs soft classification probability as the evaluation standard, leading to a more balanced data distribution across levels.

These results reflect the different emphases of the two methods regarding data quality assessment and classification capability. When comparing these methods, we conclude that while the manual filtering-based method exhibits a clear advantage in recognition capability, it may lack adaptability for specific categories. Conversely, the grading method based on model soft classification probability, despite potential shortcomings in recognition capability, offers superior handling of data uncertainty, providing a more nuanced view of data classification. Future research could consider integrating these two methods, leveraging the meticulousness of manual screening with the flexibility of probability grading, to further enhance the overall efficacy of data classification. This integrated approach can improve classification accuracy and adapt to complex datasets, increasing the model's effectiveness and reliability in real-world applications.

Model performance evaluation. Models were trained on both the complete dataset and sample datasets categorized by different quality levels as determined by two evaluation methods. Manual inspection was carried

First level category	5.0	4.0	3.0	2.0	1.0	0.0
Res	12583 (30.9%)	874 (2.1%)	1133 (2.8%)	14959 (36.7%)	1399 (3.4%)	9734 (23.9%)
Com	754 (11.30%)	246 (3.7%)	168 (2.5%)	4050 (60.6%)	32 (0.5%)	1434 (21.4%)
Pub	1792 (28.5%)	604 (9.6%)	105 (1.7%)	2618 (41.6%)	17 (0.2%)	1150 (18.3%)
Ind	9810 (40%)	1013 (4.1%)	1283 (5.2%)	4363 (17.8%)	9 (0.1%)	8020 (32.7%)
Agr	10118 (47.2%)	4701 (22%)	749 (3.5%)	428 (2%)	5 (0.1%)	5410 (25.3%)
Tra	0 (0%)	0 (0%)	0 (0%)	571 (71.4%)	4 (0.5%)	224 (28%)
Unk	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (0.1%)	25058 (99.9%)
Total	35057 (27.9%)	7438 (5.9%)	3438 (2.7%)	26989 (21.5%)	1477 (1.1%)	51030 (40.7%)

Table 3. Distribution of quality levels for each category in multi-level model classification method based on manual filtering.

First level category	5.0	4.0	3.0	2.0	1.0	0.0
Res	15769 (38.7%)	3713 (9.1%)	2834 (6.9%)	7233 (17.7%)	1399 (3.4%)	9734 (23.9%)
Com	1088 (16.3%)	420 (6.3%)	602 (9%)	3108 (46.4%)	32 (0.5%)	1434 (21.4%)
Pub	637 (10.1%)	1096 (17.4%)	741 (11.7%)	2645 (42.1%)	17 (0.2%)	1150 (18.3%)
Ind	3699 (15.1%)	7145 (29.1%)	7283 (29.7%)	3795 (15.5%)	9 (0.1%)	2567 (10.5%)
Agr	11999 (56%)	5116 (23.8%)	2090 (9.7%)	1258 (4.2%)	5 (0%)	943 (4.4%)
Tra	0 (0%)	0 (0%)	0 (0%)	571 (71.4%)	4 (0.5%)	224 (28%)
Unk	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (0.1%)	25058 (100%)
Total	33192 (26.5%)	17490 (13.9%)	13550 (10.8%)	18610 (14.8%)	1477 (1.1%)	41110 (32.8%)

Table 4. Distribution of quality levels for each category in grading method based on model soft classification probability.

out on the sample set to eliminate the influence of sample quality on the experiments. From this, A total of 5000 samples were randomly selected, and cross-validation was performed by core team members to ensure 100% accuracy with respect to the true values (1000 samples per category). This process ensures there are no obvious errors within the samples and that the consistency between functional attributes and labeling results is maintained, serving as the test set for verifying the quality of the validation dataset. The experiments in this study were conducted using the Pytorch framework, with acceleration provided by an RTX A4000 16 G GPU. For all models, the learning rate, number of iterations, and batch size were set to 0.01 decay to 0.0001, 100, and 8, respectively. The networks were trained using the Adam optimizer. The confusion matrix in Fig. 3 and Table 5 display the performance of the model utilized in this study. The deep learning model demonstrates excellent effectiveness on both the entire sample set and the dataset with a quality level of 5, achieving a testing accuracy of 0.86 and a kappa of 0.804 on the full sample dataset as shown in Fig. 3(a). The prediction accuracy is notably higher for residential, industrial, and agricultural land compared to public and commercial land. As shown in Fig. 3(b), in the level 5 dataset assessed by the multi-level model classification, the testing accuracy reaches 0.975 and the kappa is 0.965. As shown in Fig. 3(c), in the level 5 dataset evaluated by soft classification, the testing accuracy is 0.895 with a kappa of 0.832. These results underscore that models trained on high-quality small datasets outperform those trained on large datasets of lower quality, offering superior capability in identifying land use functions.

Comparing the accuracy of models trained with different quality levels, Table 5 and the confusion matrix in Fig. 4((b–k)) show the data of various grades. These grades are derived from the dataset evaluated by the multi-level model classification method. Levels 4 and 5 exhibit the highest capabilities in land use function recognition. These results affirm that higher-quality datasets, with fewer erroneous, confused, and low-quality samples, can significantly enhance the model's recognition performance. Conversely, the effectiveness of the grading method based on model soft classification probabilities in screening high-quality datasets is limited. This limitation is likely because certain samples with high probability for some categories are entirely misclassified, preventing the soft classification method from effectively distinguishing between high-quality and low-quality samples. A significant portion of the dataset consists of level 0 samples, which lack both point of POI information and explicit land use classification. We retain these samples rather than discard them, as the absence of information itself may serve as a meaningful signal. Specifically, level 0 samples often correspond to underdeveloped areas, such as rural farmland or marginal regions, where infrastructure and POI density are inherently low. In this context, the absence of POI data becomes a crucial distinguishing characteristic, aiding in differentiating underdeveloped areas from urbanized regions. This implicit feature can be utilized in spatial pattern analysis to assess the correlation between POI sparsity and factors like economic development level and infrastructure distribution.

Land use mapping results.

The distribution of MSLU-100K data across cities nationwide is depicted in Fig. 5. Due to regional differences in economic development and land use complexity across the country, particularly with a higher concentration of southern cities, the sample density is greater in the southern regions.

Model	TA	Kappa
MSLU-100K Model	0.860	0.816
Level 5(Multi-level Model)	0.975	0.965
Level 4(Multi-level Model)	0.911	0.848
Level 3(Multi-level Model)	0.797	0.702
Level 2(Multi-level Model)	0.640	0.343
Level 1(Multi-level Model)	0.562	0.106
Level 5(Soft Classification)	0.895	0.832
Level 4(Soft Classification)	0.858	0.794
Level 3(Soft Classification)	0.726	0.552
Level 2(Soft Classification)	0.670	0.416
Level 1(Soft Classification)	0.509	0.056

Table 5. Performance of different methods on different datasets.

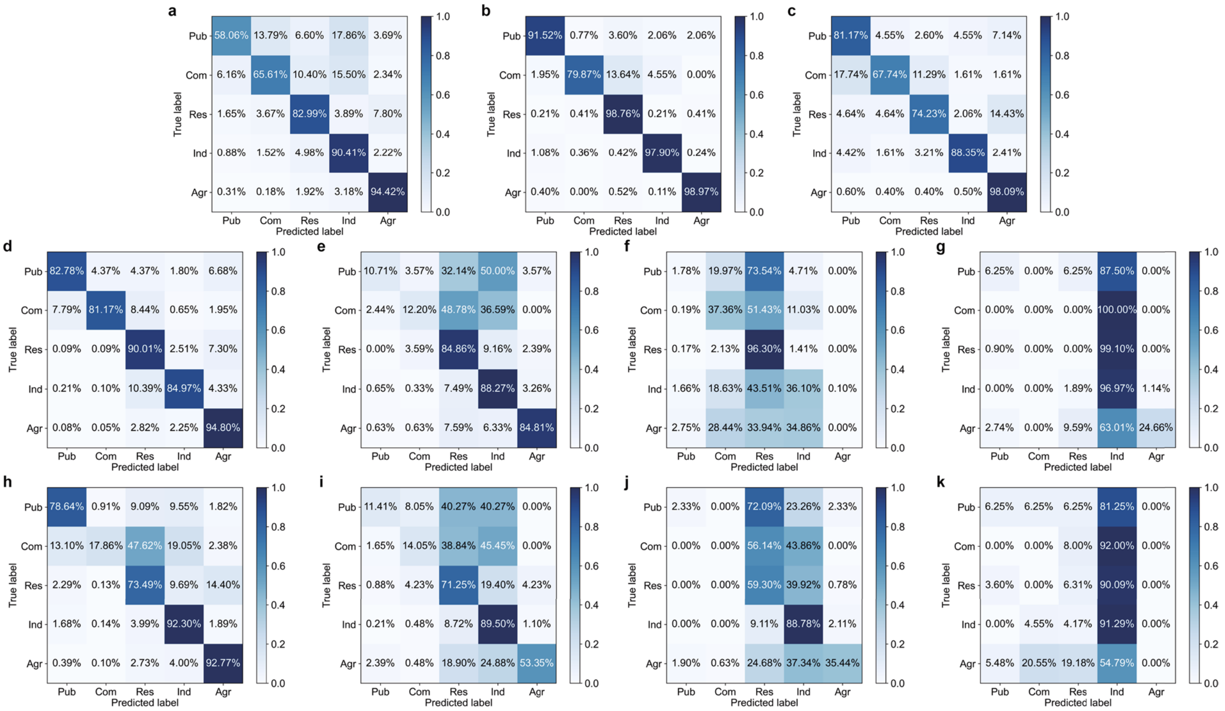

Fig. 4 Confusion matrix obtained by (**a**) MSLU-100K model, (**b**) 5-level dataset assessed by the multi-level model classification, (**c**) 5-level dataset assessed by soft classification, (**d**) 4-level dataset assessed by the multi-level model classification, (**e**) 3-level dataset assessed by the multi-level model classification, (**f**) 2-level dataset assessed by the multi-level model classification, (**g**) 1-level dataset assessed by the multi-level model classification, (**h**) 4-level dataset assessed by soft classification, (**i**) 3-level dataset assessed by soft classification, (**j**) 2-level dataset assessed by soft classification, (**k**) 1-level dataset assessed by soft classification.

Consequently, the occurrence of more dense samples in the south compared to the north is attributed to the urbanization process, complexity of land use, and economic development disparities.

To validate the usability of the dataset, models were trained using the constructed land use dataset and validated nationwide to conduct land use predictions, with the accuracy of predictions visualized. Results in Fig. 6 indicate that the prediction accuracy reached 71.5%. In addition, we performed a statistical analysis of the prediction errors for parcels of different sizes in the city depicted in Fig. 6. We define parcels smaller than 5000 square meters as small parcels and those larger than 50000 square meters as large parcels. We calculated the proportion of prediction errors for these parcels. The statistical results are detailed in Table 6. It is evident that the prediction accuracies of the large and small parcels we defined are both lower than the overall accuracy. Some large parcels are located in remote areas where POI data is sparse, making it difficult to determine land use categories. Moreover, due to the substantial area of these large parcels, they may internally encompass multiple land use types, but the model tends to classify them based on predominant features, which increases the error. In big cities, some complex large parcels may be misclassified due to mixed uses. In addition, Due to their limited area, small parcels contain fewer POI and remote sensing feature information, making it challenging to provide

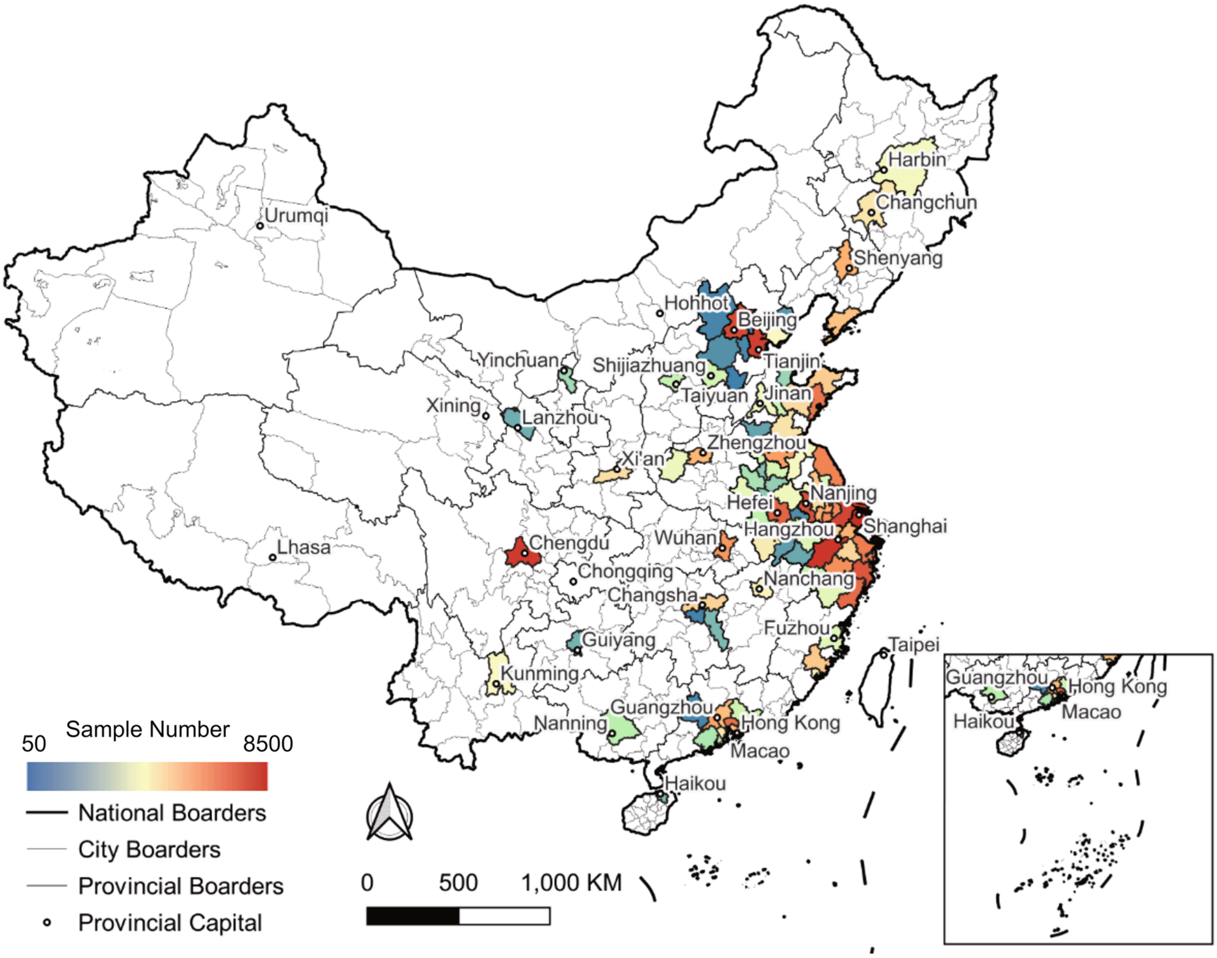

Fig. 5 Distribution of MSLU-100K dataset samples in cities across the country.

sufficient basis for discrimination. In urban dense areas, the land use categories among small parcels are highly similar, leading to confusion in the model.

Because the majority of the MSLU-100K dataset is concentrated in the Yangtze River Delta region, the model more readily captures the characteristics of this area during the learning process, leading to higher mapping accuracy there. Moreover, the abundance of external architectural features and internal socioeconomic attributes further boosts the region's performance. Conversely, due to the relative scarcity of data in northern cities and western regions, the model's learning efficacy is poorer in these areas, resulting in lower mapping accuracy. Northern cities often exhibit architectural styles and urban planning features distinct from those of southern cities, which may challenge the model's ability to effectively recognize these features with limited samples. Furthermore, the geographical environment and varying levels of development in western regions result in data scarcity, thereby impacting mapping accuracy.

Figure 7 illustrates that in developed regions with high population density, urban land use is primarily concentrated on commercial, residential, and public facility construction. In contrast, more land in relatively underdeveloped areas is allocated for agriculture or remains undeveloped. The eastern plains, conducive to large-scale urban expansion and industrial development, exhibit relatively well-planned urban construction. The western mountainous regions tend to preserve more natural reserves and ecological land[41]. Historic cities such as Xi'an and Nanjing maintain their historical heritage, influencing the pattern of urban land use[42]. Emerging cities like Shenzhen experience rapid urban expansion, and their land use more prominently displays characteristics of modernization and internationalization[43].

This study constructed the MSLU-100K multi-source land use dataset for major Chinese cities, effectively addressing significant data gaps. It proposed a data quality assessment system that combines manual methods with deep learning to precisely evaluate dataset quality. This system introduces a new methodological standard for evaluating land use dataset quality and serves as an important reference for microscale urban land use classification, as well as urban development and planning efforts.

Dataset limitations.

To address the subjectivity and labeling errors encountered in previous data labeling processes, we developed a human-computer collaboration method. This method enhances labeling quality through cross-checking, feedback, and correction of annotated data. However, these measures remain subject to

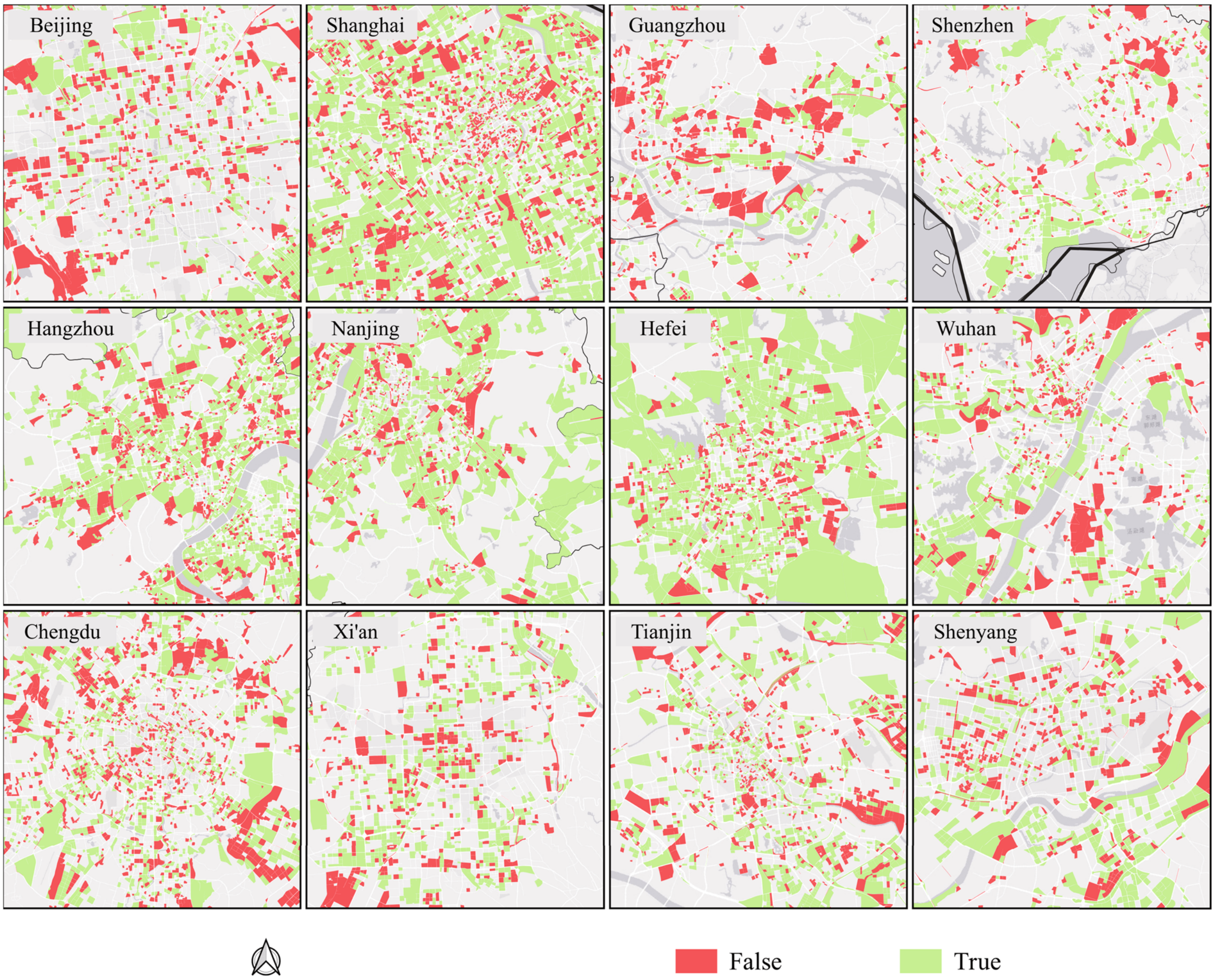

Fig. 6 Correctness judgment of land use prediction.

City	Prediction accuracy of large parcels	Prediction accuracy of small parcels
Beijing	53.8%	61.6%
Shanghai	62.5%	64.7%
Guangzhou	53.2%	63.3%
Shenzhen	62.2%	54.8%
Hangzhou	66.5%	65.3%
Nanjing	70.4%	61.5%
Hefei	69.4%	70.7%
Wuhan	60.3%	54.8%
Chengdu	53.9%	59.7%
Xi'an	60.7%	59.1%
Tianjin	64.4%	51.1%
Shenyang	53.2%	60.7%

Table 6. Prediction accuracy of parcels of different sizes.

human bias, resulting in occasional labeling errors and significant time consumption. Future research is aimed at developing applications that leverage models for auxiliary checks and real-time feedback, thereby alleviating these issues and improving labeling efficiency and accuracy. Additionally, the recognition accuracy for the "commercial area" category remains insufficient, particularly when it is confused with "industrial land". This issue is especially pronounced within the context of Chinese cities, where mixed-use commercial areas often intertwine commercial functions with other categories. Consequently, recognizing commercial land is more challenging than other land use categories. Furthermore, the definitions and boundaries of secondary categories such as "commercial area"

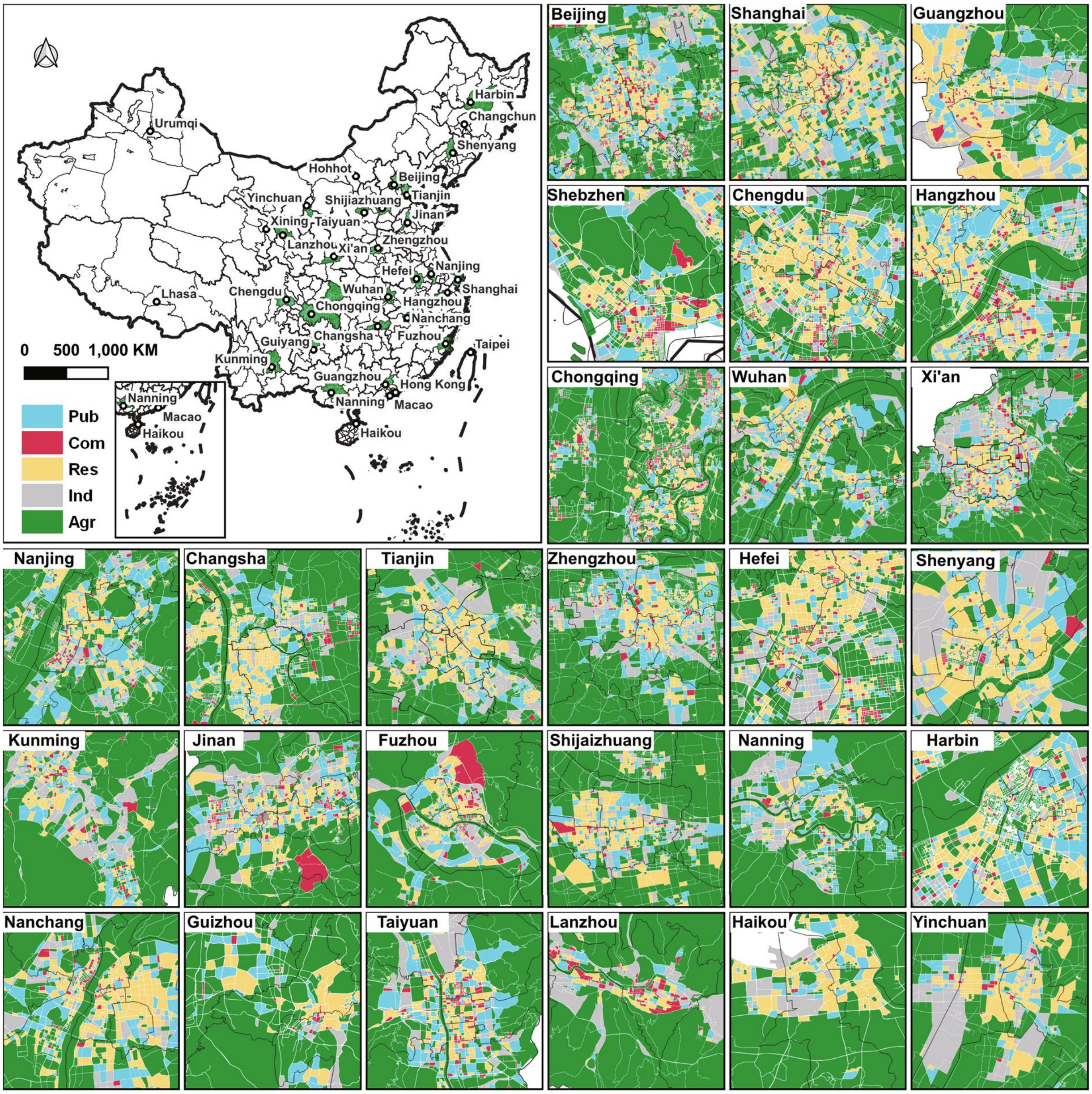

Fig. 7 Land use classification map of Chinese first-tier cities and some provincial capitals.

and "industrial land" are often ambiguous, adding to the recognition challenges. This problem also leads to low overall accuracy of the model for the land use mapping task. In future studies, we plan to overcome the current limitations by enhancing data balancing and incorporating more advanced model interpretability technologies to better manage complex land use scenarios. Additionally, we intend to explore land use classification models with enhanced accuracy and performance, using the MSLU-100K dataset. Our aim is to ensure high accuracy for both manual and model-based labeling of the dataset, and to select evaluation methods and standards wisely to further enhance classification outcomes.

Usage Notes

The MSLU-100K dataset consists of approximately 100,000 irregular remote sensing parcel images, covering 81 cities across China and encompassing all administrative levels. Utilizing the "Urban Land Classification and Planning Construction Land Standard" (GB 50137–2011) and Alibaba's Gaode Map POI, we categorized the features captured by the remote sensing images based on their primary uses into five major categories: "Residential Land", "Commercial and Service Facilities Land", "Industrial Land", "Public Management and Public Service Facilities Land", and "Agricultural and Natural", with secondary categories under each major category, resulting in a total of 22 subcategories. Additionally, during the labeling process, we identified a limited number of "Transportation Facilities Land", as well as "Unknown Land Use" categories with insufficient parcel information making determination difficult, which are also included in the dataset. Consequently, the final dataset comprises 7 major categories and 28 subcategories.

Code availability

The software used to create the dataset were an intelligent data annotation platform developed by Alibaba (https://imark.taobao.com) and Python 3.9. The rest of the code and sample data used to reproduce our work are publicly available at https://doi.org/10.6084/m9.figshare.27852591.

Received: 11 December 2024; Accepted: 22 April 2025;
Published online: 07 May 2025

References

1. Lyu, Y., Wang, M., Zou, Y. & Wu, C. Mapping trade-offs among urban fringe land use functions to accurately support spatial planning. *Science of The Total Environment* **802**, 149915 (2022).
2. Fang, Z. *et al*. Impacts of land use/land cover changes on ecosystem services in ecologically fragile regions. *Science of The Total Environment* **831**, 154967 (2022).
3. Yan, X. *et al*. A multimodal data fusion model for accurate and interpretable urban land use mapping with uncertainty analysis. *International Journal of Applied Earth Observation and Geoinformation* **129**, 103805 (2024).
4. Ouma, Y. O. *et al*. Urban land-use classification using machine learning classifiers: comparative evaluation and post-classification multi-feature fusion approach. *European Journal of Remote Sensing* **56**, 2173659 (2023).
5. Lu, W., Tao, C., Li, H., Qi, J. & Li, Y. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sensing of Environment* **270**, 112830 (2022).
6. Whang, S. E., Roh, Y., Song, H. & Lee, J.-G. Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal* **32**, 791–813 (2023).
7. Qiao, W. & Huang, X. Assessment the urbanization sustainability and its driving factors in Chinese urban agglomerations: An urban land expansion - Urban population dynamics perspective. *Journal of Cleaner Production* **449**, 141562 (2024).
8. Zhang, X., Du, S., Zhou, Y. & Xu, Y. Extracting physical urban areas of 81 major Chinese cities from high-resolution land uses. *Cities* **131**, 104061 (2022).
9. Yao, Y. *et al*. Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation* **106**, 102664 (2022).
10. Jelinski, D. E. & Wu, J. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecol* **11**, 129–140 (1996).
11. Fotheringham, A. S. & Wong, D. W. S. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environ Plan A* **23**, 1025–1044 (1991).
12. Wu, K., Wang, D., Lu, H. & Liu, G. Temporal and spatial heterogeneity of land use, urbanization, and ecosystem service value in China: A national-scale analysis. *Journal of Cleaner Production* **418**, 137911 (2023).
13. Shi, W. *et al*. Reliability and consistency assessment of land cover products at macro and local scales in typical cities. *International Journal of Digital Earth* **16**, 486–508 (2023).
14. Xiao, Y. *et al*. Class imbalance: A crucial factor affecting the performance of tea plantations mapping by machine learning. *International Journal of Applied Earth Observation and Geoinformation* **129**, 103849 (2024).
15. Zhao, S. *et al*. Land Use and Land Cover Classification Meets Deep Learning: A Review. *Sensors* **23**, 8966 (2023).
16. Martín, L., Sánchez, L., Lanza, J. & Sotres, P. Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation. *Internet of Things* **22**, 100779 (2023).
17. Guan, Q., Cheng, S., Pan, Y., Yao, Y. & Zeng, W. Sensing Mixed Urban Land-Use Patterns Using Municipal Water Consumption Time Series. *Annals of the American Association of Geographers* **111**, 68–86 (2021).
18. Zhou, W., Newsam, S., Li, C. & Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing* **145**, 197–209 (2018).
19. Cheng, G., Han, J. & Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE* **105**, 1865–1883 (2017).
20. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification | IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8736785.
21. Zhu, Q. *et al*. Knowledge-guided land pattern depiction for urban land use mapping: A case study of Chinese cities. *Remote Sensing of Environment* **272**, 112916 (2022).
22. Chen, H., Chen, J. & Ding, J. Data Evaluation and Enhancement for Quality Improvement of Machine Learning. *IEEE Trans. Rel.* **70**, 831–847 (2021).
23. Javanmard, R., Lee, J., Kim, J., Liu, L. & Diab, E. The impacts of the modifiable areal unit problem (MAUP) on social equity analysis of public transit reliability. *Journal of Transport Geography* **106**, 103500 (2023).
24. Xia, C., Yeh, A. G.-O. & Zhang, A. Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: A case study of five Chinese megacities. *Landscape and Urban Planning* **193**, 103669 (2020).
25. Zhong, B., Xing, X., Love, P., Wang, X. & Luo, H. Convolutional neural network: Deep learning-based classification of building quality problems. *Advanced Engineering Informatics* **40**, 46–57 (2019).
26. Mohammadi, P., Ebrahimi-Moghadam, A. & Shirani, S. Subjective and Objective Quality Assessment of Image: A Survey.
27. Nehmé, Y. *et al*. Textured Mesh Quality Assessment: Large-Scale Dataset and Deep Learning-based Quality Metric. Preprint at http://arxiv.org/abs/2202.02397 (2023).
28. Li, Y. Research and Application of Deep Learning in Image Recognition. in *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)* 994–999 https://doi.org/10.1109/ICPECA53709.2022.9718847 (IEEE, Shenyang, China, 2022).
29. Chen, G., Pei, Q. & Kamruzzaman, M. M. Remote sensing image quality evaluation based on deep support value learning networks. *Signal Processing: Image Communication* **83**, 115783 (2020).
30. Kang, L., Ye, P., Li, Y. & Doermann, D. Convolutional Neural Networks for No-Reference Image Quality Assessment. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1733–1740 https://doi.org/10.1109/CVPR.2014.224 (IEEE, Columbus, OH, USA, 2014).
31. Wu, H. *et al*. DCAI-CLUD: a data-centric framework for the construction of land-use datasets. *International Journal of Geographical Information Science* 1–24 https://doi.org/10.1080/13658816.2024.2387200 (2024).
32. Chen, B. *et al*. Mapping essential urban land use categories with open big data: Results for five metropolitan areas in the United States of America. *ISPRS Journal of Photogrammetry and Remote Sensing* **178**, 203–218 (2021).
33. Sinclair, R. VON THÜNEN AND URBAN SPRAWL. *Annals of the Association of American Geographers* **57**, 72–87 (1967).
34. Liu, Z. *et al*. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9992–10002 https://doi.org/10.1109/ICCV48922.2021.00986 (IEEE, Montreal, QC, Canada, 2021).
35. Srivastava, S., Vargas-Muñoz, J. E. & Tuia, D. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment* **228**, 129–143 (2019).

36. Huang, W., Cui, L., Chen, M., Zhang, D. & Yao, Y. Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science* **36**, 1905–1930 (2022).
37. Fawaz, H. I. *et al*. InceptionTime: Finding AlexNet for Time Series Classification. *Data Min Knowl Disc* **34**, 1936–1962 (2020).
38. Zhu, P. *et al*. Deep Learning for Multilabel Remote Sensing Image Annotation With Dual-Level Semantic Concepts. *IEEE Trans. Geosci. Remote Sensing* **58**, 4047–4060 (2020).
39. Liu, W., Wen, Y., Yu, Z. & Yang, M. Large-Margin Softmax Loss for Convolutional Neural Networks. Preprint at http://arxiv.org/abs/1612.02295 (2017).
40. Yao, Y. MSLU-100K: A Large Multi-Source Dataset for Land Use Analysis in Major Chinese Cities. OSF https://doi.org/10.17605/OSF.IO/YAENR (2025).
41. Jin, G. *et al*. Trade-offs in land-use competition and sustainable land development in the North China Plain. *Technological Forecasting and Social Change* **141**, 36–46 (2019).
42. Zhao, Y., Ponzini, D. & Zhang, R. The policy networks of heritage-led development in Chinese historic cities: The case of Xi'an's Big Wild Goose Pagoda area. *Habitat International* **96**, 102106 (2020).
43. Fei, W. & Zhao, S. Urban land expansion in China's six megacities from 1978 to 2015. *Science of The Total Environment* **664**, 60–71 (2019).

Acknowledgements

We are extremely grateful for the comments and suggestions provided by the editor and reviewers. This work was supported by the National Natural Science Foundation of China [42471491, 42171466], the Alibaba Group through Alibaba Innovation Research Program [20228670], the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources [KF-2023-08-04], the Guangdong-Hong Kong-Macau Joint Laboratory Program [2020B1212030009] and the "CUG Scholar" Scientific Research Funds at China University of Geosciences (Wuhan) [2022034].

Author contributions

Yao Yao - Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition; Yueheng Ma - Methodology, Software, Validation, Writing - Original Draft, Writing - Review & Editing; Ronghui Gao - Validation, Writing - Original Draft, Writing - Review & Editing; Xiaoqin Yan - Validation, Writing - Original Draft, Writing - Review & Editing; Qingfeng Guan - Writing - Review & Editing, Supervision, Funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.