

## DCAI-CLUD: a data-centric framework for the construction of land-use datasets

Hao Wu, Zhangwei Jiang, Anning Dong, Ronghui Gao, Xiaoqin Yan, Zhihui Hu, Fengling Mao, Hong Liu, Pengxuan Li, Peng Luo, Zijin Guo, Qingfeng Guan & Yao Yao

**To cite this article:** Hao Wu, Zhangwei Jiang, Anning Dong, Ronghui Gao, Xiaoqin Yan, Zhihui Hu, Fengling Mao, Hong Liu, Pengxuan Li, Peng Luo, Zijin Guo, Qingfeng Guan & Yao Yao (2024) DCAI-CLUD: a data-centric framework for the construction of land-use datasets, International Journal of Geographical Information Science, 38:11, 2379-2402, DOI: [10.1080/13658816.2024.2387200](https://doi.org/10.1080/13658816.2024.2387200)

**To link to this article:** <https://doi.org/10.1080/13658816.2024.2387200>



View supplementary material [↗](#)



Published online: 05 Aug 2024.



Submit your article to this journal [↗](#)



Article views: 213



View related articles [↗](#)



View Crossmark data [↗](#)

RESEARCH ARTICLE



## DCAI-CLUD: a data-centric framework for the construction of land-use datasets

Hao Wu<sup>a</sup> , Zhangwei Jiang<sup>b</sup> , Anning Dong<sup>a</sup> , Ronghui Gao<sup>a</sup> ,  
Xiaoqin Yan<sup>c</sup> , Zhihui Hu<sup>a</sup> , Fengling Mao<sup>b</sup> , Hong Liu<sup>b</sup> ,  
Pengxuan Li<sup>b</sup> , Peng Luo<sup>c,d</sup> , Zijin Guo<sup>a</sup> , Qingfeng Guan<sup>a</sup>  and  
Yao Yao<sup>a,e,f</sup> 

<sup>a</sup>School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei Province, P. R. China; <sup>b</sup>Alibaba Group, Hangzhou, Zhejiang Province, China; <sup>c</sup>Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China; <sup>d</sup>Chair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany; <sup>e</sup>Center for Spatial Information Science, The University of Tokyo, Chiba, Japan; <sup>f</sup>Guangdong – Hong Kong – Macau Joint Laboratory for Smart Cities, Shenzhen, China

### ABSTRACT

A high-quality land-use dataset is crucial for constructing a high-performance land-use classification model. Due to the complexity and spatial heterogeneity of land-use, the dataset construction process is inefficient and costly. This challenge affects the quality of datasets, consequently impacting the model's performance. The emerging field of Data-Centric Artificial Intelligence (DCAI) is expected to deliver techniques for dataset optimization, offering a promising solution to the problem. Therefore, this study proposes a data-centric framework named DCAI-CLUD for the construction of land-use datasets. Based on this framework, the accuracy and rate of data labeling are improved by 5.93 and 28.97%. The *Gini* index of the dataset and the proportion of samples with non-mixed land-use categories are enhanced by 3.27 and 8.52%. The overall accuracy (OA) and Kappa of the land-use classification model improved significantly by 27.87 and 58.08%. This study is the first to introduce DCAI into the field of geographic information and remote sensing and verify its effectiveness. The proposed framework can effectively improve the construction efficiency and quality of the dataset and synchronously optimize the model performance. Based on the proposed framework, we constructed a multi-source land-use dataset of major cities in China named CN-MSLU-100K.

### HIGHLIGHTS

1. A framework for optimizing the land-use dataset construction process is proposed.
2. Filtering and pre-labeling improved the quality and efficiency of data labeling.
3. The performance of land-use classification model is enhanced by dataset optimization.

### ARTICLE HISTORY

Received 5 January 2024  
Accepted 25 July 2024

### KEYWORDS

Land-use classification; data-centric artificial intelligence; point-of-interest; remote sensing image; multi-source data fusion

- 4. Preconceived results have a subjective impact on the data labelers.
- 5. The first study to introduce DCAI for land-use classification is launched.

1. Introduction

Accurate land-use classification is an essential basis for urban planning and sustainable urban development and can effectively reflect regional socio-economic contributions (Zhou *et al.* 2020) and explore the impacts of land-use change on the ecological environment (Kumar and Arya 2021). With the development of deep learning techniques and spatio-temporal geographic data, land-use classification based on machine learning models have been widely studied (Huang *et al.* 2022, Lu *et al.* 2022, Yao *et al.* 2022).

Significant time invested in constructing machine learning models is spent on preparing the training data. The quality of this data directly influences the overall performance of the model (Whang *et al.* 2023). Therefore, high-quality land-use datasets are the basis for constructing high-performance land-use classification models.

The characteristics of land-use data present challenges for dataset construction, dataset quality, and model training (Figure 1). The land-use classification includes three processes: data sampling, data labeling, and model training. With global urbanization, the size of urban areas continues to grow, and the land-use categories are becoming more and more complex (Xia *et al.* 2020, Yao *et al.* 2022, Zhang *et al.* 2022).

During the data sampling process, the modifiable areal unit problem (MAUP) (Fotheringham and Wong 1991, Jelinski and Wu 1996) arises from multi-scale land parcels, which means that the training of the model must include data of multiple scales. Furthermore, due to the spatial heterogeneity (Wu *et al.* 2023) of land-use, knowledge learned from one area is difficult to transfer directly to other regions. Therefore, the spatial distribution of the data must be considered during the sampling process.

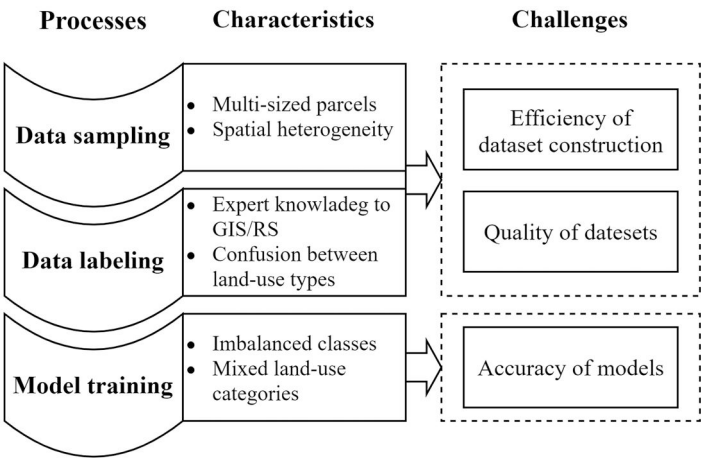


Figure 1. The processes of constructing a land-use classification model and the challenges therein due to the specificity of the land-use data.

In the data labeling process, visual interpretation of land-use requires expertise in the fields of geographical sciences. Additionally, land-use categories are numerous and many of them are easily confused with each other (Zhang *et al.* 2022). The above facts make the construction of land-use datasets based on visual interpretation or automatic labeling by artificial intelligence challenging, thus affecting the efficiency of dataset construction and the quality of the dataset.

During the model training process, there is an obvious imbalance in the categories of land-use (Wang *et al.* 2022, Zhu *et al.* 2022). The class imbalance problem poses a challenge to the training of the machine learning models (Lin *et al.* 2017, Wasikowski and Chen 2010). Moreover, mixed land-use categories are common in modern cities (Guan *et al.* 2021). Data of mixed land-use categories cannot be used in the training dataset because they cannot be given an explicit label. Therefore, the quality of land use data poses a challenge to the accuracy of land use classification models.

Given the spatial heterogeneity of land-use data, generalizing a model built on a dataset from one region to other regions is a challenge. Many studies have constructed land-use datasets. These datasets contain varying amounts of land-use categories and focus on different regions. For example, PatternNet (Zhou *et al.* 2018), NWPU-RESISC45 (Cheng *et al.* 2017), EuroSAT (Helber *et al.* 2019), and ILU- CUG (Zhu *et al.* 2022). However, it is necessary to create new datasets when constructing land-use classification models in different regions.

Existing studies on land-use classification describe few details of dataset labeling. Existing data annotation methods can be categorized into three types: manual, automatic and semi-automatic. Manual annotation methods are time-consuming and labor-intensive. Several studies have introduced a method that combines automatic model labeling, effectively assisting in data labeling (Maihami and Yaghmaee 2018, Zhu *et al.* 2020). However, the correctness of labels automatically generated by the model depends on the performance of the model, so ensuring the quality of the datasets is challenging. Semi-automated data labeling can strike a balance between data acquisition efficiency and data quality (Zhu *et al.* 2020). For example, using machine learning models to make predictions on unlabeled data (also called pre-labeling) is used to assist in manual labeling. However, further study is still needed on how to adopt a semi-automatic labeling approach to systematically optimize the labeling process and improve the efficiency and quality of the dataset.

Data-centric Artificial Intelligence (DCAI), an emerging concept in the field of AI, has received extensive attention from both academia and industry (Polyzotis and Zaharia 2021, Jakubik *et al.* 2024). Compared to Model-Centric Artificial Intelligence (MCAI), DCAI emphasizes the centrality of data in AI systems (Hamid 2022, Zha *et al.* 2023). A large part of the machine learning process is spent on data preparation. Without high-quality data, even the best machine-learning models cannot perform well (Whang *et al.* 2023).

Optimizing data quality includes acquiring new data and optimizing existing data. For instance, Motamedi *et al.* (2021) optimized the existing dataset by data cleaning, label checking, and addressing the class imbalance problem. They generated new data using Generative Adversarial Networks (GANs), leading to a 5% improvement in model accuracy. Zhong *et al.* (2022) enhanced the robustness of the model by incorporating

transferable adversarial examples and 14 kinds of common corruptions into the dataset. Lin *et al.* (2022b) proposed RoboFlow, which orchestrates the development pipelines of AI-enhanced robots. With the integration of new data, the robot system can be updated swiftly and efficiently. In summary, DCAI, serving as a guideline for AI model construction, offers a solution to the issue of land-use dataset construction. However, due to the complexity and specificity of land-use classification studies, applying DCAI for land-use dataset construction still necessitates further study and practice.

To address the above problems, this study optimized the construction process of the land-use dataset based on the guiding principle of DCAI. We proposed a data-centric framework for the construction of land-use datasets called DCAI-CLUD. In this study, an irregular parcel-scale land-use dataset CN-MSLU-100K was constructed based on DCAI-CLUD. This study verified the effectiveness of the proposed method in enhancing the quality of datasets and performance of models from three perspectives: data labeling efficiency, dataset quality, and the accuracy of the model.

2. Methodology

The study process consists of three main parts (Figure 2): (a) The implementation methodology of DCAI-CLUD, which includes sample filtering based on the location and size of parcels, and a ‘human-computer collaboration’ approach to dataset construction. (b) The evaluation of the effectiveness of DCAI-CLUD is based on evaluation indices from three perspectives: data labeling, dataset, and model. (c) This study

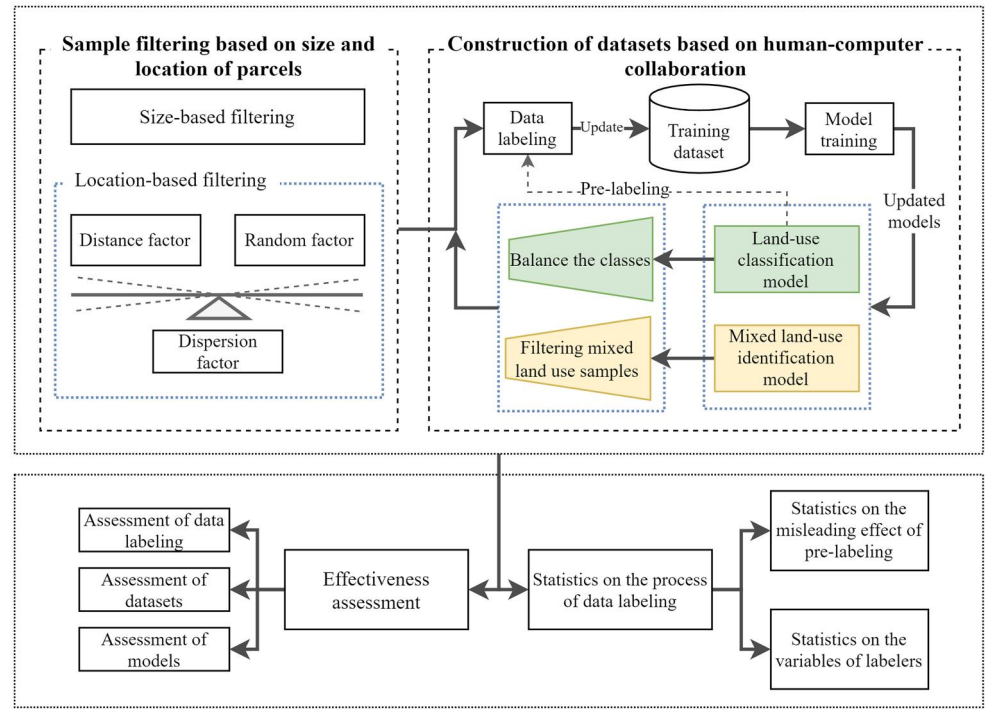


Figure 2. The workflow of this study. It includes the proposed DCAI-CLUD framework, the assessment of the effectiveness of DCAI-CLUD, and the statistics of the data labeling process.

analyzed the performance of labelers in the labeling process, aiming to provide insights into the possible implications of the proposed framework.

### 2.1. Sample filtering based on size and location of parcels

The land-use data used in this study consists of irregular parcels generated by road networks. Due to overlapping road network data, some parcels are too small in area, while some parcels are excessively large due to insufficient road network data. These parcels are considered noise in model training because they cannot be effectively input into the model. Moreover, according to von Thünen's 'land rent theory', the size of a parcel affects its land-use category by influencing the expected profit (Sinclair 1967). Therefore, we statistically evaluated the variation of the evaluation indices (Section 2.3) for the quality of the dataset with respect to the size of the parcels. Then, based on the result, an attempt was made to find an appropriate size range to filter out the noisy data and reduce the degree of class imbalance in the dataset and the proportion of parcels with mixed land-use categories.

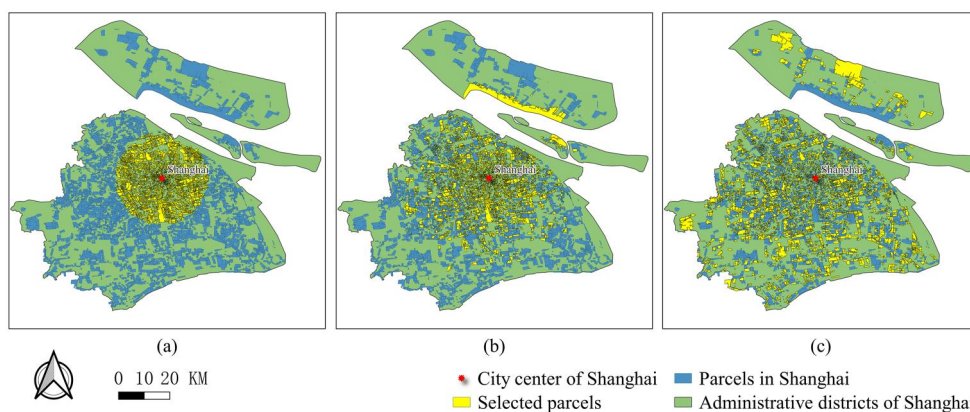
Land-use in urban areas is more complex and diverse than that in non-urban areas, but the degree of land-use mixture has increased along with urban growth (Guan *et al.* 2021). To obtain more diverse land-use data in urban areas while simultaneously maximizing the quality of the dataset, this study proposed a sample filtering method based on the location of parcels. This method mediates the degree of dispersion of sampled parcels from the city center to the city periphery using a dispersion factor  $d$ . Using this method, this study examined the relationship between data quality and  $d$  and then determined the optimal value of  $d$ . The proposed method is as follows:

Calculate the probability  $B_i^p$  that a parcel  $B_i$  within a city is selected by using Equation (1). Herein,  $B_i^r$  is a uniformly distributed random value between 0 and 1, which we call the random factor corresponding to  $B_i$ .  $B_i^r$  is used to represent the random sampling method. To increase the probability that a parcel in the city center area will be selected,  $B_i^d$ , the distance from  $B_i$  to the city center, is used to divide  $B_i^r$ . We refer to  $B_i^d$  as the distance factor.  $N(B_i^d)$  is the normalization result of  $B_i^d$ , serving to eliminate the influence of the dimensions of  $B_i^d$ , and is obtained using Equation (2).  $d$  is the dispersion factor, representing the degree to which the sampling results are influenced by the random factor  $B_i^r$ .

The value of  $d$  lies between 0 and 1. The larger the value of  $d$ , the more spatially dispersed the selected parcels become. When  $d$  equals 1,  $B_i^p = 1/B_i^d$  and the selected parcels are clustered in the city center (Figure 3(a)). When  $d$  equals 0.5,  $B_i^r$  and  $B_i^d$  have equal weights, the selected parcels are equally influenced by both factors (Figure 3(b)). When  $d$  equals 0,  $B_i^p = B_i^r$ . At this point the sampling is random and the selected parcels are evenly distributed (Figure 3(c)).

$$B_i^p = \frac{(B_i^r)^d}{(N(B_i^d))^{(1-d)}} \quad (1)$$

$$N(x) = \frac{x - \text{Min}(x)}{\text{Max}(x) - \text{Min}(x)} \quad (2)$$



**Figure 3.** The effect of the parcel selection when setting different dispersion factors, as an example in Shanghai. (a)  $d=0$ . (b)  $d=0.5$ . (c)  $d=1$ .

After the above steps, a total of  $n$  parcels in the current city are arranged in descending order based on the selection probability, forming the set  $P = [p_1, p_2, \dots, p_n]$ . From the set of parcels  $P$ , the first  $k$  are selected based on demand.

## 2.2. Construction of datasets based on human-computer collaboration

Assisting manual labeling with the predictive power of machine models is an effective way to improve labeling efficiency (Zhu *et al.* 2020). This study proposed a ‘human-machine collaboration’ approach to dataset construction based on model pre-labelling. By continuously updating the data, the land-use classification model was iterated, and each iteration was used for the pre-labeling of the next round of data labeling. The subsequent round of labeling was then purposefully filtered based on the pre-labeling results, facilitating a collaborative iteration of the dataset and the model.

The proposed method is as follows: (a) Obtain the initial dataset  $D_0$  by manual labeling without pre-labeling. Then train the land-use classification model  $M_0^l$  and the mixed-category sample identification model  $M_0^m$  based on  $D_0$ . (b) Predict the remaining samples to be labeled ( $D_r$ ) using  $M_0^l$  and  $M_0^m$  to get the pre-labeled class  $l$  and the label  $m$  of whether it is a mixed land use for each sample. Store the prediction results in the fields of the dataset to be labeled. (c) Starting from the second round of labeling, the remaining amount demanded in each class is  $r = [r_1, r_2, \dots, r_N]^T$ , which is obtained by subtracting the cumulative labeled amount of the class from the total demanded amount of each class. (d) Based on the label  $m$ , the samples identified as mixed land use are filtered out from  $D_r$ . The remaining data to be labeled after filtering constitutes the dataset  $D_r'$ . (e) Calculate the number of samples to be selected in each class according to Equation (3), where  $s = [s_1, s_2, \dots, s_N]^T$  is the number of parcels selected from  $D_r'$  for each class.  $p_x^y$  is the probability that the true class is  $x$  and is predicted by the model to be  $y$ , and is obtained through the confusion matrix of  $M_0^l$ .  $r$  is the number of requirements for each category. (f) After calculating the number of samples for each category in  $s$ , a random sample of each class is drawn from  $D_r'$  according to the label  $l$  to obtain the data to be labeled in the next round.



(g) Perform the next round of labeling to obtain the dataset  $D_i$  ( $i = 1, 2, 3, \dots$ ). Then iterate the models  $M_i^l$  and  $M_i^m$  using  $D_i$ . (h) Return to step (a) and continue iterating the above process until the model  $M_n^l$  satisfying the accuracy requirement is obtained.

$$s = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} p_1^1 & p_1^2 & \cdots & p_1^N \\ p_2^1 & p_2^2 & \cdots & p_2^N \\ \vdots & \vdots & \ddots & \vdots \\ p_N^1 & p_N^2 & \cdots & p_N^N \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} = P^{-1}r \quad (3)$$

In step (g) above, the incremental dataset is used for a new round of model training. In order to avoid overfitting of the model, this study uses the newly acquired data to fine-tune the model or retrain it using the full amount of data. The model with the highest accuracy based on each round of dataset was obtained through tuning. The model was continued to be used for the next round of data labeling.

In the above method, the classification model used in DCAI-CLUD was BDF-Net, which is a two-branch neural network coupling remote sensing images and POI data (Figure S1). It consists of two branches: one based on the Transformer (Vaswani *et al.* 2017) for extracting features from remote sensing images and the other using POI embedding (Yao *et al.* 2017) for extracting features from POI data. Then, calculates the weights of the two features through an adaptive feature weighting layer (Lu *et al.* 2022).

To prove the effectiveness of BDF-Net, an ablation analysis of the model was performed, as shown in Table S1. A total of six comparative models were constructed, including the ablation of the data used in the model, the sampling method used for remote sensing images, and the Transformer model used for remote sensing image feature extraction. Then, 10,000 samples were randomly selected from the CN-MSLU-100K dataset, which was constructed in this study, to train the models. The accuracy and confusion matrices of the six models are shown in Table S1 and Figure S2. The results demonstrate the effectiveness of each module of the models. The OA of BDF-Net is 0.881, and the Kappa coefficient is 0.878.

Note that BDF-Net was chosen as an example model to embed DCAI-CLUD in this study, but any classification model can be used to pre-label the data. This is since the performance of any common model will improve with the quality of the data (Whang *et al.* 2023).

### 2.3. Effectiveness assessment of DCAI-CLUD

Several datasets were constructed using the above method, and models were trained using these datasets. A blank control group was established for comparison, that is, a dataset was constructed and a model was trained without using any optimization methods. To assess the quality of the datasets and Models, the following evaluation indices were established:

The efficiency of dataset construction is evaluated by the accuracy of the labeling result Acc (Equation (4)) and the rate of labeling Rate (Equation (5)). During the manual labeling process, the labelers are asked to sample check each other to ensure the



accuracy of the labeling results. In Equation (6),  $M$  represents the number of samples inspected by each of the  $n$  labelers, while  $C$  represents the number of samples with correct inspection results.  $Rate$  is the number of labels each labeler makes per hour. In formula (5),  $N$  represents the number of samples labeled by each of the  $n$  labelers, and  $H$  represents the length of time (in hours) labeled by this labeler. The higher values of  $Acc$  and  $Rate$  represent the higher efficiency of dataset construction.

$$Acc = \frac{\sum_1^n C/M}{n} \quad (4)$$

$$Rate = \frac{\sum_1^n N/H}{n} \quad (5)$$

The quality of the dataset is assessed by three indicators: (a) The degree of class imbalance within the sample is assessed by the *Gini* impurity, which is calculated using Equation (6). In Equation (6),  $n_i$  represents the number of samples in class  $i$ ,  $N$  represents the number of all samples, and  $K$  denotes the number of classes. The *Gini* assumes a value within the range of  $[0, (K-1)/K]$ . A larger value of *Gini* signifies a more balanced distribution of classes in the dataset. (b) The percentage of samples with non-mixed land-use categories,  $P_{nm}$ , is employed to assess the percentage of samples with unambiguous labels. (c) The percentage of parcels in urban areas,  $P_{iu}$ , is employed to assess the percentage of samples of interest. To comprehensively assess the impact of *Gini*,  $P_{nm}$ , and  $P_{iu}$  on data quality, this study normalizes these three indicators to eliminate the effect of magnitude, then calculates the mean to obtain the composite score  $S_{avg}$ .

$$Gini = 1 - \sum_{i=1}^K \left( \frac{n_i}{N} \right)^2 \quad (6)$$

For the evaluation indices of the model, Overall Accuracy (OA) (Equation (7)), Kappa (Equation (8)), and Confusion Matrix are utilized. In Equations (7) and (8),  $x_{ij}$  represents the element of row  $i$  and column  $j$  of the confusion matrix,  $x_{ii}$  denotes the number of correctly predicted samples for each category, and  $N$  is the number of test samples. During model training, the dataset is divided into training and validation sets in the ratio of 7:3. The model is first trained using the training dataset, and then the validation set is used to cross-validate the accuracy of the model.

$$Overall\ Accuracy = \frac{\sum_{i=1}^K x_{ii}}{N} \quad (7)$$

$$Kappa = \frac{\sum_{i=1}^K x_{ii}/N - \sum_{i=1}^K x_{ii} \left( \sum_{j=1}^K x_{ij} \sum_{j=1}^K x_{ji} \right) / N^2}{1 - \sum_{i=1}^K x_{ii} \left( \sum_{j=1}^K x_{ij} \sum_{j=1}^K x_{ji} \right) / N^2} \quad (8)$$

## 2.4. Statistical methods for the dataset construction process

When using models to assist labelers in semi-automatic annotation, the impact of labelers' subjective behavior is an important issue that cannot be ignored. To answer the question of 'whether model pre-labeling will mislead the labelers?' we conducted

a comparative experiment both with and without pre-labeling. Initially, the labelers were permitted to label with pre-labeling. Then, according to the time interval of the Ebbinghaus memory curve (Ebbinghaus 2013), after an interval of more than one month, the labelers were requested to re-label their previously labeled samples without pre-labeling, without being informed of this change. By comparing the outcomes of the two labeling processes, the probability of being influenced by pre-labeling was calculated.

Managing the labelers is a crucial task to enhance the efficiency of data labeling. During the labeling process, three variables for each labeler were calculated and recorded: the number of labeled samples (*Num*), the working hours (*WH*), and the labeling rate (*Rate*). The correlation between these three variables and the accuracy of their labeling (*Acc*) was subsequently measured using the Pearson correlation coefficient (Equation (9)) (Cohen *et al.* 2009). The results were used to analyze the behavioral characteristics of each labeler and any potential patterns that may be embedded in the labeling process. In Equation (9), the Pearson correlation coefficient is derived by calculating the covariance (*cov*) as well as the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of each of the two variables.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \quad (9)$$

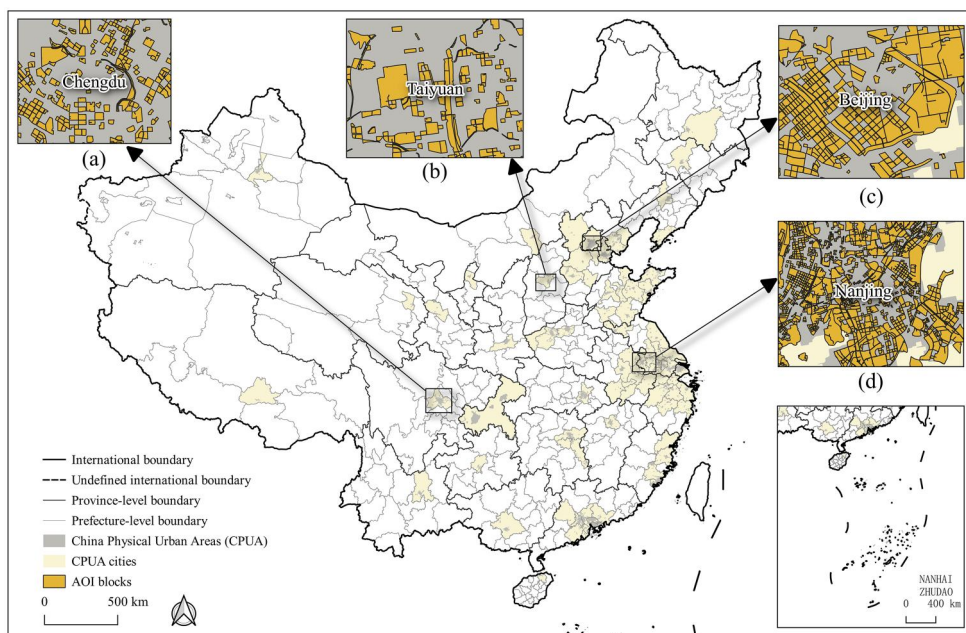
### 3. Results

#### 3.1. Study areas and data

Eighty-one cities in China were used as the study area to construct the land-use dataset. These cities cover all administrative levels in China and possess diverse spatial forms, patterns, and landscapes (Zhang *et al.* 2022), enabling the dataset to better represent Chinese urban areas. The total area of the study area is 983,215 km<sup>2</sup>, as shown in Figure 4. A list of the 81 cities with their administrative levels is shown in Table S2.

The unlabeled parcel boundary data, also called the Area of Interest (AOI) data, was obtained from Alibaba, the largest e-commerce company in China, and was generated using geometric algorithms based on the road network data. This study also used the China Physical Urban Area (CPUA) data produced by Zhang *et al.* (2022) to calculate the percentage of labeled parcels within the urban area. Figure 4 shows the preview of the CPUA and AOI data in the study area for the four representative cities. Among them, Beijing is the capital city and an important political, economic and cultural center of China, which is a representative city reflecting the development of urbanization in China. Chengdu, Taiyuan, Beijing and Nanjing are located in the southwest, north-central and southeast of China, respectively. They are distributed in different directions in China, which can well reflect the urban landscape and spatial characteristics of different regions in China.

The Remote Sensing Imagery (RSI) data was downloaded from Google Earth Engine (GEE). These RSIs were created by fusing and stitching together data from multiple sources, including Landsat, Quick Bird, IKONOS, SPOT5, and aerial photography.



**Figure 4.** The China Physical Urban Area (CPUA) and Area of Interest (AOI) data in the study area encompass 81 major cities in China. Four representative examples, including (a) Chengdu, (b) Taiyuan, (c) Beijing, and (d) Nanjing, are illustrated. (It is IJGIS policy to remain strictly neutral with respect to jurisdictional claims on disputed territories in published maps, and the naming conventions used in maps are left to the discretion of authors.).

The downloaded RSIs contain three bands (red, green, blue), produced from 2019 to 2022, with a resolution of 2.5 m.

The POI dataset was mainly obtained through the application program interface (API) provided by the Gaode Open Platform (<https://lbs.amap.com/>). Gaode Maps is a leading provider of digital map content, navigation, and location service solutions in China. On the basis of Gaode POIs, combined with Alibaba POI database, we finally obtained 80 million pieces of data, including 32 primary classifications and 170 secondary classifications. The POI data are proxies for real-world locations that can effectively reflect the functional structure of the city (Psyllidis *et al.* 2022). The POI data was used to assist the RSI in the land-use classification.

The land-use of the parcels was categorized into five major first-level categories and 22 second-level categories (Table S5). A group of 56 labelers was enlisted to perform the labeling. During the labeling process, labelers were asked to sample and cross-check 25% of the data with each other. For tasks with less than 90% accuracy, the labelers were asked to revise their results until the accuracy exceeded 90%. Ultimately, a total of about 100,000 pieces of data were obtained, including 40,682 pieces of residential districts (Res), 6,286 pieces of public services land (Pub), 6,684 pieces of commercial zones (Com), 24,498 pieces of industrial land (Ind), and 21,411 pieces of agricultural and natural land (Agr). We named the final training dataset CN-MSLU-100K (China Multi-Source Land-use Dataset). A demo dataset named CN-MSLU-DEMO has been publicly available. Refer to the supporting material for a detailed description of the dataset.

### 3.2. Results of sample filtering based on location and size of parcels

Before proceeding with more labeling, this study first experimented with the proposed sample filtering method using a total of 31,960 samples from nine cities, which include four first-tier cities and five new first-tier cities in China, covering the east, west, south, north, and center of the country. The cities are Beijing, Shanghai, Guangzhou, Shenzhen, Hangzhou, Xi'an, Changsha, Tianjin, and Wuhan. Then, the parcels were equally divided into 20 groups according to their size. Refer to Table S3 for detailed grouping data.

The statistical results (Figure 5(a~b)) show that when the area of the parcels is less than 38,931.32 m<sup>2</sup> (Group 1), more than 50% of them fall into 'agricultural and natural land'. The *Gini* index (0.621) and  $P_{iu}$  (0.835) are lower, but  $P_{nm}$  (0.954) is higher compared to parcels with larger area.

When the area ranges from 38,931.32 m<sup>2</sup> to 676,818.47 m<sup>2</sup> (Groups 2–16), compared to smaller parcels, the percentage of 'agricultural and natural land' decreases rapidly by 30%, while the samples of other land-use categories begin to increase. The *Gini* index,  $P_{nm}$ , and  $P_{iu}$  change by 20.82, -7.47, and 13.72%. Within this interval, the *Gini* index and  $P_{nm}$  tend to stabilize, with the fluctuations accounting for 12.28 and 10.84% of the total fluctuations.  $P_{iu}$  rapidly decreases by -95.78% (from 0.742 to 0.031) until it is close to 0 after the area exceeds 340,493.105 m<sup>2</sup> (Groups 14–20), indicating that the parcels are gradually shifting from urban to rural areas.

When the area exceeds 676,818.47 m<sup>2</sup> (Groups 17–20), urban facilities such as 'public services land' and 'industrial land' are decreasing, while 'residential districts',



**Figure 5.** Statistics of *Gini*,  $P_{nm}$ ,  $P_{iu}$  and land-use categories. (a) Variation of *Gini*,  $P_{nm}$ ,  $P_{iu}$  with area. (b) The change of land-use percentage with area. (c) Variation of *Gini*,  $P_{nm}$ ,  $P_{iu}$  with dispersion factors  $d$ . (d) The change of land-use percentage with dispersion factors  $d$ .

which are primarily composed of large-scale rural homesteads, as well as ‘agricultural and natural land’ are increasing. This result means the parcels are gradually located far away from the city. The *Gini* and  $P_{nm}$  change by  $-27.61\%$  and  $10.72\%$ . In addition, we found that when the area of parcels is less than  $38,931.32\text{ m}^2$  and more than  $676,818.47\text{ m}^2$ , the Pearson correlation coefficient between *Gini* and  $P_{nm}$  is  $-0.914$ , showing a very strong negative correlation.

For the proposed sample filtering method based on the location of parcels (refer to Table S4 for detailed grouping data), the results in Figure 5(c~d) show that the location of selected parcels gradually moves away from the city center, and the various categories of land-use change gently and uniformly as  $d$  increases from 0 to 1. Among them, the proportions of ‘agricultural and natural land’ and ‘industrial land’ increase by 39.88 and 61.10%. The proportion of ‘commercial zones’, ‘residential districts’, and ‘public services land’ decreases by 23.58, 20.04, and 21.06%. The *Gini* index and  $P_{nm}$  fluctuate and increase, while  $P_{lu}$  decreases at an accelerating rate. Finally,  $S_{avg}$  presents a fluctuating increase, achieving a maximum value of 0.752 at  $d = 0.8$ .

Based on the above results, dataset  $D_s$  was acquired by limiting the size of the parcels on the original dataset  $D_{ori}$  which was purely manually labeled, from 38931.35 to  $676,818.47\text{ m}^2$ . The models  $M_{ori}$  and  $M_s$  were then trained using the two datasets. The evaluation results of the dataset and model are presented in Table 1. The results indicate that, compared to  $D_{ori}$ , the *Gini* and  $P_{lu}$  of  $D_s$  increase by 1.09 and 19.26%. When compared to  $M_{ori}$ , the OA and Kappa of  $M_s$  increase by 4.92 and 6.15%.

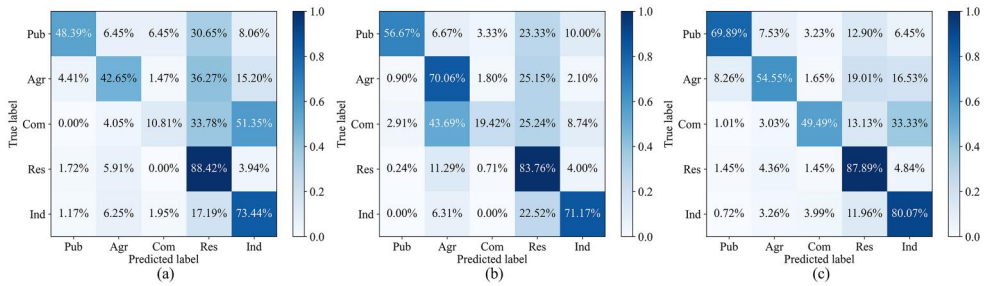
The confusion matrix (Figure 6) reveals that due to the imbalance in the classes of the samples, the classification accuracy of  $M_{ori}$  for different classes also appears to be extremely imbalanced. Given that ‘residential land’ has the largest number of samples, the model tends to predict more samples as ‘residential land’ to ensure accuracy. In comparison to  $M_{ori}$ ,  $M_{sl}$ , which has samples more concentrated in urban areas and has filtered out fine and very large parcels, enhances the recognition accuracy of ‘public services land’, ‘agricultural and natural land’, and ‘commercial zones’ by 17.11, 64.27, and 79.65%. The accuracy of ‘agricultural and natural land’ shows the most significant improvement. Meanwhile, the recognition accuracy of ‘residential districts’ and ‘industrial land’ is slightly reduced by 5.27 and 3.09%.

Further, based on  $D_{sl}$ , the sample  $D_{sl}$  was obtained by filtering based on spatial location, taking  $d$  equal to 0.8. The model  $M_{sl}$  was then trained. The results (Table 1) demonstrate that, in comparison to  $D_{sl}$ ,  $D_{sl}$  increases the *Gini*,  $P_{nm}$ , and  $P_{lu}$  by 2.97, 0.38, and 10.65%. In comparison to  $D_{ori}$ , these metrics exhibit changes of +4.09,  $-1.38$ , and

**Table 1.** Results of evaluation indices for datasets constructed with different sample filtering methods and models trained with the datasets.

Dataset	Results of dataset construction			Results of model training	
	<i>Gini</i>	$P_{um}$	$P_{lu}$	OA	Kappa
$D_{ori}$	0.733	0.798	0.701	0.671	0.520
$D_s$	0.741	0.784	0.836	0.704	0.552
	(+1.09%)	( $-1.75\%$ )	(+19.26%)	(+4.92%)	(+6.15%)
$D_{sl}$	0.763	0.787	0.925	0.762	0.664
	(+2.97%)	(+0.38%)	(+10.65%)	(+8.24%)	(+20.29%)

$D_{ori}$ : a completely randomly sampled dataset.  $D_s$ : a dataset filtered by size-based method.  $D_{sl}$ : a dataset filtered by size-location-based method.



**Figure 6.** Confusion matrix of the models trained on datasets constructed with different sample filtering methods. (a) Model  $M_{ori}$  trained on a completely randomly sampled dataset. (b) Model  $M_s$  trained on a dataset filtered by size-based method. (c) Model  $M_{sl}$  trained on a dataset filtered by size-location-based method.

**Table 2.** Results of evaluation indices for datasets constructed by different labeling methods and models trained with the datasets.

Dataset	Data construction efficiency		Quality of datasets		Results of model training	
	Acc	Rate	Gini	$P_{nm}$	OA	Kappa
$D_{np}$	89.92%	116.34	0.690	0.749	0.778	0.649
$D_p$	90.00%	115.94	0.779	0.733	0.739	0.662
	(+0.09%)	(−0.34%)	(+12.90%)	(−2.14%)	(−5.01%)	(+2.00%)
$D_{pm}$	95.25%	150.04	0.757	0.866	0.858	0.822
	(+5.83%)	(+29.41%)	(−2.82%)	(+18.14%)	(+16.10%)	(+24.17%)

$D_{np}$ : a dataset whose labeling was done without pre-labeling.  $D_p$ : a dataset whose labeling process used pre-labeling.  $D_{pm}$ : a dataset whose labeling process used pre-labeled and filtered mixed land use samples.

+31.95%. In terms of model training, the OA and Kappa of  $M_{sl}$  rise by 8.24 and 20.29% when compared to  $M_s$ , ultimately displaying a total increase of 13.56 and 27.69% when compared to  $M_{ori}$ .

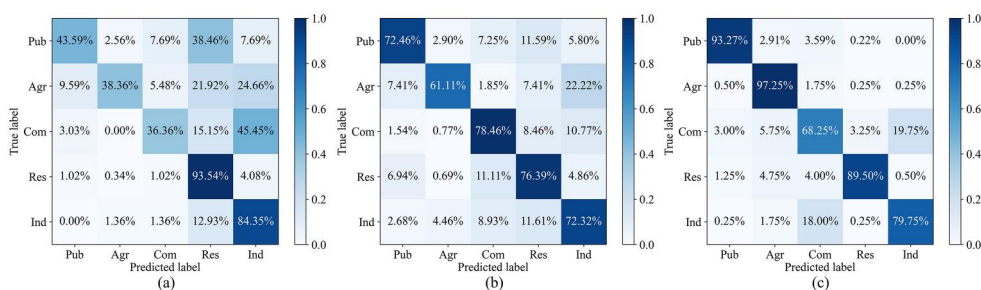
The confusion matrices (Figure 6) reveal that, in comparison to  $M_s$ , due to the more balanced classes of the dataset and the increased number of samples from urban areas, the recognition accuracy of  $M_{sl}$  for ‘public services land’, ‘commercial zones’, ‘residential districts’, and ‘industrial land’ has improved by 23.33, 154.84, 4.93, and 12.51%. The most significant improvement is observed in the accuracy of ‘commercial zones’. However, the recognition accuracy of ‘agricultural and natural land’ has declined by 22.14%.

### 3.3. Results of dataset construction based on human-computer collaboration

Labelers were organized to label the data both without and with pre-labeling during the same period to avoid the effect caused by the proficiency difference. Based on this method, the study acquired datasets  $D_{np}$  and  $D_p$  and used them to train models  $M_{np}$  and  $M_p$ .

Table 2 indicates that  $D_p$  improves its Gini by 12.90% compared to  $D_{np}$  due to the reduction of class imbalance in the dataset by pre-labeling. Although  $M_p$ ’s OA decreases by 5.01%, its Kappa improves by 2.00% compared to  $M_{np}$ . The confusion matrix results (Figure 7) reveal that the accuracy of  $M_{np}$  for each class appears unbalanced due to the class imbalance of the sample.  $M_{np}$ ’s recognition accuracy for ‘public





**Figure 7.** Confusion matrix for models trained on datasets constructed based on human-computer collaboration. (a) Models  $M_{np}$  trained on a dataset whose labeling is done without pre-labeling. (b) Model  $M_p$  trained on a dataset whose labeling process used pre-labeling. (c) Models  $M_{pm}$  trained on a dataset whose labeling process used pre-labeling and filtered samples with mixed land-use categories.

services land', 'agricultural and natural land' and 'commercial zones' are all below 44%. In contrast, as the classes of the dataset become more balanced,  $M_p$ 's recognition accuracies for all three categories improve to over 60%.

Building on the pre-labeling of the samples, the proposed method of filtering samples with mixed categories was further applied to the samples to derive the dataset  $D_{pm}$  on which  $M_{pm}$  was trained. Due to the filtering of samples with mixed categories, which are relatively more difficult for labelers to recognize, the *Acc*, *Rate* and  $P_{nm}$  of  $D_{pm}$  improve by 5.83, 29.41 and 18.14%, compared to  $D_p$ . In comparison to  $M_p$ ,  $M_{pm}$ 's OA and Kappa improve by 16.10 and 24.17%. The confusion matrix results (Figure 7) reveal that, compared to  $M_p$ , the recognition accuracy of  $M_{pm}$  for 'public services land', 'agricultural and natural land', 'residential districts', and 'industrial land' increase by 28.72, 59.14, 17.16, and 10.27%. Among them, the accuracies of 'public services land' and 'agricultural and natural land' exceed 90%. However, the precision of 'commercial zones' drops by 13.01, with 19.75% of the samples misclassified as 'industrial land'.

When comparing  $D_{pm}$  and  $D_{np}$ , it is found that *Acc*, *Rate*, *Gini*, and  $P_{nm}$  improve by 5.93, 28.97, 9.71, and 15.62%. When  $M_{pm}$  is compared with  $M_{np}$ , it is found that OA and Kappa improve by 10.28 and 26.66%. When  $D_{pm}$  is compared with the completely randomly constructed dataset  $D_{ori}$ , the *Gini* and  $P_{nm}$  are found to improve by 3.27% (0.733–0.757) and 8.52% (0.798–0.866). Lastly, when comparing  $M_{pm}$  with  $M_{ori}$ , OA and Kappa are found to significantly improve by 27.87% (0.671–0.858) and 58.08% (0.520–0.822).

### 3.4. Statistical results and analysis of the dataset construction process

To analyze whether the pre-labeling misled the labelers, we selected 1197 samples (evenly covering five categories) to conduct the above experiment. The labelers were organized to label the same samples with and without pre-labeling, with a one-month interval between the two labeling tasks. The results indicate that the error rate is 24.5% when pre-labeling is applied. Among these errors, 7.10% are attributed to the labelers being misled by pre-labeling, accounting for 28.98% of all error causes.

Typical labeling errors are illustrated in Figure 8. Among them, 87.06% of the errors occur when the labelers are misled by the pre-labeling and choose the wrong





**Figure 8.** Typical examples of labeling errors. Errors that occur due to the misleading of pre-labeling include (a) industrial land being labeled as public services land and (b) mixed land-use being labeled as commercial zones. Errors that occur due to the confusion in second-level land-use categories include (c) ‘villas and high-end residences’ or ‘high-rise residential buildings’ and (d) ‘urban village’ or ‘rural homestead’.

first-level class. For instance, in [Figure 8\(a\)](#), a hospital under construction should be categorized as ‘industrial land’. However, it was labeled as ‘public services land’, aligning with the pre-labeling label. In [Figure 8\(b\)](#), a parcel that contains about 50% ‘residential districts’ and 50% ‘commercial zones’ should be labeled as mixed land-use. However, it was labeled as ‘commercial zones’.

The remaining 12.94% of the errors occurred when the first-level class was labeled correctly, but the second-level class was mislabeled. Part of the reason for this error is that in this study, the pre-labeling only predicted the first-level class of the parcels, and the second-level class was randomly selected. This approach may have misled the labelers. This kind of error typically occurs between conceptually confusing land-use categories. For example, [Figure 8\(c\)](#) depicts a residential neighborhood. However, it remains to be seen whether this parcel is a ‘villas and high-end residences’ because there are villa-like buildings in it. In [Figure 8\(d\)](#), it is easy for labelers to identify this parcel as a ‘residential district’. However, due to the rapid development of cities in China, there are many areas in the transition stage between urban and rural areas (Lang *et al.* 2016), making it relatively challenging to distinguish whether it belongs to an ‘urban village’ or a ‘rural homestead’.

[Table 3](#) shows the value of Pearson’s correlation coefficients between the number of labeled samples (*Num*), the working hours (*WH*), the rate of labeling (*Rate*) for all labelers, and their labeling accuracy (*Acc*). In this study, a Pearson correlation coefficient between 0 and 0.4 is considered a weak correlation, between 0.4 and 0.8 is considered a moderate correlation, and above 0.8 is considered a strong correlation.

The results indicate that the Pearson’s correlation coefficients between *Num*, *WH*, and *Acc<sub>i</sub>* are primarily concentrated in the range of 0 to 0.4 (62.06% for *Num* and 70.68% for *WH*). However, 31.03 and 22.4% exhibit moderate positive correlation with *Acc*. In contrast, the correlation between *Rate* and *Acc* is more pronounced. 49.99% of

**Table 3.** The value of Pearson's correlation coefficients between the number of labeled samples (*Num*), the working hours (*WH*), the rate of labeling (*Rate*) for all labelers, and their labeling accuracy (*Acc*).

	<i>Num</i>	<i>WH</i>	<i>Rate</i>
The value of Pearson's correlation coefficient	Percentage of persons (%)		
−1~−0.8	0.00	0.00	1.72
−0.8~−0.6	0.00	1.72	1.72
−0.6~−0.4	3.44	0.00	3.44
−0.4~−0.2	1.72	3.44	3.44
−0.2~0	1.72	1.72	1.72
0~0.2	18.96	13.79	15.51
0.2~0.4	43.10	56.89	15.51
0.4~0.6	25.86	15.51	32.75
0.6~0.8	5.17	6.89	17.24
0.8~1	0.00	0.0	6.89

labelers demonstrate a moderate positive correlation, and 6.89% show a strong correlation. It is also observed that 5.15% of labelers exhibit a moderate negative correlation, and 1.72% display a strong negative correlation.

## 4. Discussion

### 4.1. The effectiveness of sample filtering based on size and location of parcels

Filtering data based on parcel size and location can effectively improve dataset quality and model performance. The proposed method solves the problem of the lack of effective data filtering methods before data labeling and can guide researchers in making a preliminary screening of samples through sample statistics before starting data labeling, thereby obtaining higher-quality samples to be labeled and significantly improving the efficiency of subsequent labeling work. Compared to the unfiltered dataset, the model trained on the dataset obtained using the proposed filtering method improved the OA and Kappa by 13.56 and 27.69%.

The proposed sample filtering method can help researchers understand the characteristics of various categories of parcels in terms of their size and location so that land-use data can be filtered in a more precise and targeted way. Our conclusions can provide an important reference for land-use researchers and urban planning departments.

In the process of analyzing the size and location of land parcels, we discovered that as the size of the parcel changes, its land-use category and location also exhibit different characteristics. This conclusion is consistent with the findings of von Thünen's 'land rent theory' (Sinclair 1967). As the parcel area increases, the categories of parcels gradually change from urban facilities to rural and then to natural wilderness, and the location of the parcel gradually moves away from the city center. Furthermore, parcels with mixed land-use categories are mainly concentrated in the urban center area, further confirming the study of Guan *et al.* (2021). By understanding the characteristics of various categories of parcels in terms of area and spatial location, researchers can filter the land-use data more precisely and in a more targeted manner.

This study finds a new pattern for the relationship between the size and land-use categories of parcels. Our sample filtering method, based on these findings, enhances

the dataset quality and model performance without leading to a lack of certain data categories. We found that extremely small or large parcels see a rise in mixed land-use categories, leading to a more balanced category distribution. The conclusion is that the Pearson correlation coefficient between *Gini* and  $P_{nm}$  is  $-0.914$  when parcels are smaller than  $38,931.32 \text{ m}^2$  and larger than  $676,818.47 \text{ m}^2$ . Smaller parcels are mostly 'agricultural and natural lands', while larger ones are dominated by 'residential districts' and 'agricultural and natural lands'. Therefore, as parcels of certain categories dominate, the balance of categories gradually decreases, but the number of parcels with mixed land-use categories also gradually decreases. This suggests a challenge in balancing dataset classes and obtaining non-mixed samples for extreme parcel sizes. In addition, parcels that are too small or too large are not ideal samples for model training, so they were filtered out from the dataset.

We also find that the quality of the sample is optimal when the spatial distribution of the selected parcels presents a weight ratio of 8 to 2 for 'randomly dispersed' and 'concentrated around the city center'. This parameter provides a valuable reference for sample selection in land-use datasets. For the sample filtering method based on the location of parcels, the results indicate that when the dispersion factor  $d=0.8$ , it can better balance the needs of 'keeping the classes of dataset balanced', 'obtaining more non-mixed samples', and 'located in the urban area'.

#### **4.2. The effectiveness of DCAI for the construction of the land-use dataset**

This study marks the first introduction of DCAI into the field of GIS and RS, effectively applying it to land-use classification studies. The proposed DCAI-CLUD framework can significantly enhance data labeling efficiency, dataset quality, and model performance, thereby realizing the data-centric construction of land-use datasets and model training processes.

The proposed method provides an effective solution for land-use data labeling. The method addresses the issues of inefficiency and lack of guidance in labeling land-use datasets and is also applicable to other multi-category data labeling tasks. The proposed human-computer collaborative data labeling method is realized by model pre-labeling. Through model pre-labeling, priori knowledge can be gained about the unlabeled data.

Pre-labeling can improve the labeling efficiency. Based on the proposed method, the average accuracy and labeling rate of the labelers are improved by 5.93 and 28.97%. Such results are because each labeler is assigned to label only the data predicted to be the same category. This strategy can significantly reduce the learning cost of labelers and increase their proficiency in a specific category. Moreover, by filtering out some of the samples with mixed land-use categories, the redundant information in the labeling process is reduced, and the labeling rate is improved. Given that in the field of remote sensing, many datasets are required to enhance the recognition ability of the models (Tong *et al.* 2020), the methodology proposed in this study aids in generating datasets quickly and in large quantities.

Moreover, Pre-labeling can be used to equalize categories and filter unavailable data, thus reducing time wastage. Based on the proposed method, the *Gini* index of

the dataset is improved by 9.71%, and the proportion of parcels with non-mixed categories is improved by 15.62%. Ultimately, the OA and Kappa of the obtained model improved by 10.28 and 26.66%.

Using pre-labeling to assist manual labeling can ensure the accuracy of the data. Although automatically labeling data through models can improve the labeling efficiency (Maihami and Yaghmaee 2018, Zhu *et al.* 2020, Bortoloti *et al.* 2022). But the labeling process lacks rigorous inspection. The accuracy of the data is easily affected by the performance of the model. In this study, while introducing model pre-labeling to improve the labeling efficiency, manual inspection is used to effectively ensure the quality of the dataset.

This study provides a complete application scheme for data labeling based on the principles of DCAI. Our results further inspire related studies to emphasize the importance of data for geographic modeling. A core strategy of DCAI is to continuously optimize the quality of the dataset while keeping the model unchanged, thereby improving the performance of the model (Jakubik *et al.* 2024). However, when applying these generalized principles to geoscience, domain expertise is essential, and various challenges must be addressed. To handle imbalanced land-use data and mixed-category samples, our framework selects the samples to be labeled through model pre-labeling. We fixed the model structure unchanged and continuously iterated the land-use classification model and the mixed-category sample identification model by continuously improving the quality of data labeling. The model obtained from each iteration is then used to pre-label the data in the subsequent round to assist in acquiring higher-quality data. This iterative process simultaneously improves model performance and dataset quality, forming a virtuous cycle for data-centric training.

#### **4.3. Analysis of the statistical results of the data labeling process**

Investigating the potential effects of the labeling strategy on the labeler can aid in further optimizing the labeling strategy. The results remind researchers that when conducting volunteer experiments, it is important not to overlook the fact that preconceived results can have a subjective effect on volunteers. In this study, we found that 28.98% of the incorrectly labeled samples are due to being misled by pre-labeling. The likely reason for these phenomena is that labelers tend to accept pre-labeling results when the definition of rules is not sufficiently clear. As has been shown in existing research, errors in annotation are most likely since the annotator does not have a good understanding of the task in hand (Theodosiou and Tsapatsoulis 2020). To address these issues, we prompt the labelers to cross-check promptly, which can assist them in quickly correcting rules with biased understanding, ensuring the final accuracy of the dataset.

Through personalized management of different types of labelers, we can better analyze the different characteristics of the labelers, maximize their respective advantages, and enhance the overall labeling efficiency and quality. The study by Martin-Morato and Mesaros (2023) suggests that quantifying the reliability of labels by assessing annotators' capabilities can effectively improve the quality of annotated data. Therefore, it is

essential to conduct a quantitative analysis of annotators' characteristics. In this study, we find that there is a significant correlation between the rate of labeling and the accuracy of labeling. Some labelers (6.68% strong and 49.99% moderate correlation) improve in speed and accuracy over time, embodying the 'Practice makes perfect' principle. Conversely, a small group (1.72% strong and 5.16% moderate correlation) achieves better accuracy at slower speeds, reflecting the 'Slow and steady wins the race' approach. These labelers may require more time to ensure quality. Our conclusions can provide insightful references for the management and decision-making of managers in charge of data production tasks.

#### 4.4. Limitations and future works

During the data labeling process, the influence of human subjectivity is inevitable. Even though we strive to cross-check, provide feedback, and revise the labeled data, errors still occur, and a considerable amount of time is consumed. In future studies, the development of applications that utilize models to provide auxiliary checks and real-time feedback may mitigate this issue. In addition, the use of crowdsourced labeling, where labels are obtained by multi-person voting, can also help to improve the problem of labeling accuracy (Martin-Morato & Mesáros 2023, Lin *et al.* 2022a, Zhang *et al.* 2018).

According to Figure 7(c), the precision of 'commercial zones' is still not very high. And 'commercial zones' is easily confused with 'industrial land'. This is because Chinese cities have many mixed-use business districts, where commercial zones are often mixed with other categories. It is more difficult to recognize commercial land use than other land use classes (Srivastava *et al.* 2019, Yao *et al.* 2022, Yan *et al.* 2024). Moreover, the concepts of some of the secondary categories of 'commercial zones' and 'industrial land' are easily confused. However, the purpose of this study is not to propose a high-performance land use classification model, but an efficient framework for land use dataset construction. In future studies, we will explore land use classification models with higher accuracy and performance based on the DCAI-CLUD framework and the CN-MSLU-100K dataset.

It is evident that embedding a high-performance classification model in DCAI-CLUD is highly necessary. A low-performance model may impact the effectiveness of DCAI-CLUD. In future studies, we will continue to explore the effectiveness of DCAI-CLUD when embedding classification models with different performances. In addition, statistically different optimal values of the dispersion factors for different areas can further improve the quality of sampling by location. Furthermore, incorporating a mechanism to identify simple and difficult samples during the labeling process is also a potential direction for optimization in future studies. The aim of this approach is to obtain more difficult samples and thus enhance the generalization performance of models.

## 5. Conclusion

To solve the issues of low efficiency and data quality in the construction of land-use datasets, which in turn hinders the improvement of the model's performance, this study proposes a data-centric dataset construction framework named DCAI-CLUD.

As the first study introducing DCAI into a land-use classification study, the framework filters out low-quality samples based on parcels' size and location, optimizing both the dataset and the model through human-machine collaboration. Compared to the baseline method, the quality of the dataset constructed using DCAI-CLUD has improved, and the OA and Kappa of the model have also significantly increased by 27.87% and 58.08%, respectively.

The first introduction of DCAI into geographic modeling research represents a pioneering intellectual contribution to GIScience. Researchers will be inspired by the results of this study, showing that optimizing datasets is a practical and worthwhile approach to improving the usability of geographic models and that more analysis and optimization methods for geographic datasets should be further investigated in the future.

The results of this study are promising for researchers and practitioners in the field of geographic information. The proposed method helps them quickly obtain many high-quality samples in a short period and reduces the cost of dataset construction. Additionally, due to the high quality of the dataset, geographic models can achieve better performance in the initial training phase. Moreover, the conclusions derived from the analysis of labelers' behavior characteristics can assist administrators in improving the labeling process. Finally, the CN-MSLU-100K dataset constructed in this study provides a valuable land-use data resource for the field of geographic science.

Despite the impressive results of this study, the influence of human subjectivity in the data labeling process is still unavoidable. In our future work, we will develop applications that utilize the model to assist in checking and providing real-time feedback to address this issue. In addition, a low-performance model may impact the effectiveness of DCAI-CLUD. We will explore the effectiveness of DCAI-CLUD when embedding classification models with different performances in the future. And we will further incorporate a mechanism to identify simple and difficult samples during the labeling process to introduce more difficult samples to the dataset, thus helping to improve the generalization of the model.

## Acknowledgement

We are deeply grateful to Professor Yuan May, Dr. Andreas Züfle, and the anonymous reviewers for their constructive comments and suggestions on our paper. We also extend our sincere thanks to the young volunteers who contributed significantly to this human-computer collaboration project.

## Disclosure statement

No conflict of interest exists in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part.

## Funding

This work was supported by the National Key Research and Development Program of China [2023YFB3906803], the Alibaba Group through Alibaba Innovation Research Program [No. 20228670], the National Natural Science Foundation of China [42171466]; the 'CUG Scholar'



Scientific Research Funds at China University of Geosciences (Wuhan) [2022034] and a Guangdong-Hong Kong-Macau Joint Laboratory Program [2020B1212030009].

## Notes on contributors

**Hao Wu** has obtained his master's degree from China University of Geosciences (Wuhan). He is currently working at the State Grid Corporation of China. His research interests are geospatial big data mining, data-centric urban modeling. He contributed to the methodology, software developing, writing – original draft, visualization, writing – review and editing.

**Zhangwei Jiang** is a staff algorithm engineer at Alibaba Group. His research interests are LBS data mining and research & recommendation algorithm. He contributed to the project administration, conceptualization, data curation, investigation, methodology, writing – original draft, writing – review and editing.

**Anning Dong** has obtained his master's degree from China University of Geosciences (Wuhan). He is currently working at the State Administration of Foreign Exchange in China. His research interests are spatiotemporal big data mining and crime geography. He contributed to the methodology, data curation, software developing, validation, writing – original draft, writing – review and editing.

**Ronghui Gao** is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining, Interpretability of urban models. He contributed to the methodology, validation, writing – original draft, writing – review and editing.

**Xiaoqin Yan** is currently a Ph.D. student in GIScience at the Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing. His research interests are spatiotemporal big data computing and social sensing. He contributed to the methodology, data curation, validation, writing – original draft, writing – review and editing.

**Zhihui Hu** is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining, land use classification and trajectory representation learning. He contributed to the methodology, validation, writing – original draft, writing – review and editing.

**Fengling Mao** is an algorithm engineer at Alibaba Group. Her research interests are trajectory pattern mining and spatiotemporal data embedding. She contributed to the methodology, data curation, validation, software developing, writing – review and editing.

**Hong Liu** is a senior staff algorithm engineer at Alibaba Group. His research interests are data mining and research&recommendation algorithm. He contributed to the conceptualization, investigation, methodology, writing – review and editing.

**Pengxuan Li** is a senior staff data engineer at Alibaba Group. His research interests are data mining and data science. the methodology, validation, software developing, writing – review and editing.

**Peng Luo** has obtained his Ph.D. from the Chair of Cartography and Visual Analytics at the Technical University of Munich, Germany. He is about to join the Senseable City Lab at the Massachusetts Institute of Technology. His research interests include spatial association modeling, social sensing, and applied artificial intelligence. He contributed to the validation, writing – original draft, writing – review and editing.


**Zijin Guo** has obtained his master's degree from China University of Geosciences (Wuhan). He is currently working at the Changjiang Water Resources Commission in China. His research interests are trajectory data mining and complex network analysis. He contributed to the validation, writing – original draft, writing – review and editing.



**Qingfeng Guan** is a professor at China University of Geosciences (Wuhan). His research interests are high-performance spatial intelligence computation and urban computing. He contributed to the supervision, writing – review and editing.

**Yao Yao** is a Professor at China University of Geosciences (Wuhan) and a researcher at the University of Tokyo. His research interests are geospatial big data mining, analysis, and computational urban science. He contributed to the supervision, project administration, conceptualization, data curation, investigation, methodology, writing – original draft, visualization, writing – review and editing.

## ORCID

Hao Wu  <http://orcid.org/0009-0008-2767-6058>  
 Zhangwei Jiang  <http://orcid.org/0009-0002-3251-6506>  
 Anning Dong  <http://orcid.org/0009-0002-4160-5148>  
 Ronghui Gao  <http://orcid.org/0009-0003-5008-2702>  
 Xiaoqin Yan  <http://orcid.org/0009-0000-3664-1465>  
 Zihui Hu  <http://orcid.org/0009-0004-5521-189X>  
 Fengling Mao  <http://orcid.org/0009-0000-5162-5535>  
 Hong Liu  <http://orcid.org/0000-0002-8790-7478>  
 Pengxuan Li  <http://orcid.org/0009-0009-6005-6122>  
 Peng Luo  <http://orcid.org/0000-0002-3680-8509>  
 Zijin Guo  <http://orcid.org/0009-0005-4121-4341>  
 Qingfeng Guan  <http://orcid.org/0000-0002-7392-3709>  
 Yao Yao  <http://orcid.org/0000-0002-2830-0377>

## Data and codes availability statement

The CN-MSLU-100K dataset cannot be shared publicly due to the copyright reasons. However, readers can access the dataset upon request. The CN-MSLU-DEMO dataset are publicly available at <http://doi.org/10.6084/m9.figshare.24942510>. We have already provided a website for full data application and download at <https://urbancomp.net/s/cn-mslu-100k-land-use-classification-dataset-at-block-scale-for-multi-source-spatio-temporal-dataen>. The ‘human-computer collaborative’ data annotation method mentioned in Section 2.2 is operated based on a data annotation platform of Alibaba, so the code of the platform cannot be disclosed due to the copyright issues. The rest of the code and sample data used to reproduce our work are publicly available at <http://doi.org/10.6084/m9.figshare.24942510>.

## References

- Bortoloti, F.D., *et al.*, 2022. An annotated image database of building facades categorized into land uses for object detection using deep learning. *Machine Vision and Applications*, 33 (5), 80.
- Cheng, G., Han, J., and Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105 (10), 1865–1883.
- Cohen, I., *et al.*, 2009. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, 1–4.
- Ebbinghaus, H., 2013. Memory: a contribution to experimental psychology. *Annals of Neurosciences*, 20 (4), 155–156.
- Fotheringham, A.S., and Wong, D.W.S., 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A: Economy and Space*, 23 (7), 1025–1044.
- Guan, Q., *et al.*, 2021. Sensing mixed urban land-use patterns using municipal water consumption time series. *Annals of the American Association of Geographers*, 111 (1), 68–86.

- Hamid, O.H., 2022. From model-centric to data-centric AI: a paradigm shift or rather a complementary approach? In: 2022 8th International Conference on Information Technology Trends (ITT), Dubai, United Arab Emirates, 196–199.
- Helber, P., et al., 2019. Eurosat: a novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12 (7), 2217–2226.
- Huang, W., et al., 2022. Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science*, 36 (10), 1905–1930.
- Jakubik, J., et al., 2024. Data-centric artificial intelligence. *Business & Information Systems Engineering*, 1–9.
- Jelinski, D.E., and Wu, J., 1996. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, 11 (3), 129–140.
- Kumar, S., and Arya, S., 2021. Change detection techniques for land cover change analysis using spatial datasets: a review. *Remote Sensing in Earth Systems Sciences*, 4 (3), 172–185.
- Lang, W., Chen, T., and Li, X., 2016. A new style of urbanization in China: transformation of urban rural communities. *Habitat International*, 55, 1–9.
- Lin, J., Yu, T., and Wang, Z.J., 2022a. Rethinking crowdsourcing annotation: Partial annotation with salient labels for multilabel aerial image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–12.
- Lin, Q., et al., 2022b. RoboFlow: a data-centric workflow management system for developing AI-enhanced Robots. In: *Conference on Robot Learning*. PMLR, 1789–1794.
- Lin, T.-Y., et al., 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*, Paris, France, 2980–2988.
- Lu, W., et al., 2022. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sensing of Environment*, 270, 112830.
- Maihami, V., and Yaghmaee, F., 2018. Automatic image annotation using community detection in neighbor images. *Physica A: Statistical Mechanics and Its Applications*, 507, 123–132.
- Martin-Morato, I., and Mesaros, A., 2023. Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 902–914.
- Motamedi, M., Sakharnykh, N., and Kaldewey, T., 2021. A data-centric approach for training deep neural networks with less data. *arXiv preprint arXiv:2110.03613*.
- Polyzotis, N., and Zaharia, M., 2021. What can data-centric AI learn from data and ML engineering? *arXiv preprint arXiv:2112.06439*.
- Psyllidis, A., et al., 2022. Points of Interest (POI): a commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2 (1), 20.
- Sinclair, R., 1967. Von Thünen and urban sprawl. *Annals of the Association of American Geographers*, 57 (1), 72–87.
- Srivastava, S., Vargas-Muñoz, J.E., and Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 228, 129–143.
- Theodosiou, Z., and Tsapatsoulis, N., 2020. Image annotation: the effects of content, lexicon and annotation method. *International Journal of Multimedia Information Retrieval*, 9 (3), 191–203.
- Tong, X.-Y., et al., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237, 111322.
- Vaswani, A., et al., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, L., et al., 2022. Evaluation of a deep-learning model for multispectral remote sensing of land use and crop classification. *The Crop Journal*, 10 (5), 1435–1451.
- Wasikowski, M., and Chen, X.-w., 2010. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22 (10), 1388–1400.
- Whang, S.E., et al., 2023. Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, 32 (4), 791–813.

- Wu, K., et al., 2023. Temporal and spatial heterogeneity of land use, urbanization, and ecosystem service value in China: a national-scale analysis. *Journal of Cleaner Production*, 418, 137911.
- Xia, C., Yeh, A.G.-O., and Zhang, A., 2020. 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.
- Yan, X., et al., 2024. 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.
- Yao, Y., et al., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, 31 (4), 825–848.
- Yao, Y., et al., 2022. 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.
- Zha, D., et al., 2023. Data-centric AI: perspectives and challenges. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), 945–948.
- Zhang, J., et al., 2018. Improving crowdsourced label quality using noise correction. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (5), 1675–1688.
- Zhang, X., et al., 2022. Extracting physical urban areas of 81 major Chinese cities from high-resolution land uses. *Cities*, 131, 104061.
- Zhong, Y., et al., 2022. Exploiting the potential of datasets: a data-centric approach for model robustness. *arXiv preprint arXiv:2203.05323*.
- Zhou, W., et al., 2018. PatternNet: a benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, 197–209.
- Zhou, Y., Li, X., and Liu, Y., 2020. Land use change and driving factors in rural China during the period 1995–2015. *Land Use Policy*, 99, 105048.
- Zhu, P., et al., 2020. Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (6), 4047–4060.
- Zhu, Q., et al., 2022. Knowledge-guided land pattern depiction for urban land use mapping: a case study of Chinese cities. *Remote Sensing of Environment*, 272, 112916.