

Spatial assessment of settlement consolidation potential: Insights from Zhejiang Province, China

Qiushi Zhou^{1,2}, Wenze Yue^{1*}, Mengmeng Li³, Hongwei Hu¹, Leyi Zhang⁴

¹ Department of Land Management, Zhejiang University, 310058 Hangzhou, China

² School of Earth and Environmental Sciences, Cardiff University, CF10 3AT Cardiff, UK

³ School of Environmental and Geographical Sciences, Shanghai Normal University, 200234 Shanghai, China

⁴ Zhejiang Institute of Territorial Spatial Planning, 310012 Hangzhou, China

* Correspondence to: wzyue@zju.edu.cn (Prof. Wenze Yue)

E-mail address:

zhouqshi@zju.edu.cn (Qiushi Zhou)

wzyue@zju.edu.cn (Wenze Yue)

mengbjfu@126.com (Mengmeng Li)

hongwei.hu@zju.edu.cn (Hongwei Hu)

673214700@qq.com (Leyi Zhang)

This is an accepted version of manuscript to be published in
Humanities & Social Sciences Communications

Abstract: Global food security is increasingly threatened by cropland loss and rising food demand, necessitating policy interventions such as settlement consolidation. However, scaling up local consolidation practices require large-scale assessments of consolidation potential to effectively coordinate interventions for localized implementation. Here, we evaluate settlement consolidation potential across rural land systems in Zhejiang Province, by leveraging machine learning models fed with reference data from completed consolidation projects and associated explanatory variables. This consolidation potential assessment demonstrates high robustness and precision, with a recall of 0.75, a modified F1 score of 4.05, and a low coefficient of variation. The consolidation potential map illustrates heterogeneity of settlement consolidation potential across space, which identifies 29,784 hectares potential consolidation area in rural land systems, predominantly concentrated in northern prefectures such as Jiaxing (12,961 hectares) and Huzhou (8987 hectares). In general, smaller settlement patches exhibit higher consolidation potential. In addition, a distinct distribution of consolidation potential is observed across plains, hills, and mountains. Specifically, plains account for almost 80% potential consolidation area, followed by mountains (17%) and hills (3%). This study reveals the spatial heterogeneity of settlement consolidation potential, offering insights for strategies that enhance food production, optimize land use, and inform policy-making.

Keywords: Consolidation potential, Machine-learning, Positive-unlabeled learning, Food security

1. Introduction

Settlement consolidation plays a critical role in offsetting cropland loss. Historical and expected cropland losses to human settlements undermine the achievement of Sustainable Development Goals (Christopher et al. 2017; van Vliet 2019), including Zero Hunger (SDG 2) and Sustainable Cities and Communities (SDG 11). However, rural depopulation worldwide suggests a window of opportunity to counteract this loss by implementing settlement consolidation, a process that typically involves reclaiming low-density settlements into cropland (Janus and Markuszewska 2019; Daskalova and Kamp 2023).

Settlement consolidation has been intensively adopted by many countries (e.g., China, Germany, Poland). Empirical studies demonstrate that such initiatives not only increase cropland area, but also facilitate large-scale farming operations and reduce the risks associated with food deficit (Jürgenson 2016; Janus and Markuszewska 2019; Dong et al. 2022). Among these, China is the first large economy to formulate a nationwide planning for settlement consolidation, known as the “*National Land Consolidation Planning (2014-2020)*”, to address rising food demand and diminishing cropland quantities (Liu et al. 2014). This strategy is further reinforced by the “*National Land Planning of China (2016-2030)*”, which envisions the continuation of long-term consolidation efforts (NLPC 2017). Current consolidation practices reveal that the costs and benefits of settlement consolidation exhibits considerable spatial heterogeneity (Jin et al. 2017; Zang et al. 2021), influenced by a wide range of biophysical and social-economic factors (e.g., elevation, temperature, and population density). Therefore, to guarantee effective consolidation, there is an urgent need to delineate the potential for settlement consolidation at large scales that allows for interventions tailored to local contexts.

To effectively upscale local consolidation solutions to address regional challenges (Lambin et al. 2020), a fundamental step is to assess the distribution of consolidation potential, despite challenges posed by current assessment methods. While scholars have employed indoor interviews or multi-criteria evaluation models to assess consolidation potential (Gao et al. 2022a; Tao et al. 2024), existing methods meet significant challenges when applied at larger spatial scales. Specifically, interviews are typically confined to small-scale studies or individual projects due to time and financial investments

required (Cay and Uyan 2013; Colombo and Perujo-Villanueva 2019; Pan et al. 2023). In contrast, multi-criteria evaluation models, which rely on subjective weighting of various assessment dimensions, are constrained by their inherent dependence on expert knowledge, thereby limiting their broader applicability (Lin et al. 2020; Zhao et al. 2024a). These limitations underscore the need for robust, scalable methodologies that can objectively and efficiently evaluate consolidation potential across diverse spatial contexts.

Machine learning methods, particularly supervised classification approaches, have been widely applied to map consolidation potential (Xu et al. 2019; Yang et al. 2024). However, these models fall short when it comes to feeding them with merely single-labeled references, as conventional classification models require both positive and negative samples. In our case, references of suitable areas for consolidation (positive samples) are sufficient, but definitive references for areas unsuitable for consolidation (negative samples) are absent. A potential solution is to treat all unconsolidated areas (unlabeled samples) as negative samples during model training. However, this approach introduces significant biases into outputs due to the hidden presence of settlements with consolidation potential within the unlabeled dataset. This leads to class confusion—where positive samples are misclassified as negative—and feature misrepresentation, reducing the model's predictive accuracy. To address this challenge, an alternative approach is to build models with bootstrapped training datasets by combining all positive samples and a down-sampled subset of unlabeled ones with replacement (Bepler et al. 2019; McDonald et al. 2021). This method, namely Positive-Unlabeled (PU) bagging, has proved effective in dealing with the aforementioned issue in fields such as remote sensing and biological sciences (Lu and Wang 2021; Liu et al. 2024). By leveraging PU bagging, it becomes possible to produce robust and reliable evaluations of consolidation potential, particularly when combined with comprehensive datasets of consolidated settlement samples from diverse physical and socioeconomic contexts.

Ground-sourced data is crucial for conducting independent and bottom-up assessment of settlement consolidation potential. Although such data can be manually collected for individual villages through visual interpretation (Liu et al. 2023; Zhao et al. 2024b), its large-scale availability is severely hindered

by time and labor constraints. Furthermore, existing time-series datasets on land use are not suitable for acquiring reference data, as they are based on the assumption that human settlements are permanent and cannot be legally or practically converted into other land uses (Yang and Huang 2021). As a result, these datasets fail to illustrate the nuanced conversion of settlements into cropland. The absence of reliable ground-sourced data impedes the analysis of empirical relationship between the spatial distribution of consolidated settlements and explanatory variables, hindering the identification of settlements with consolidation potential. However, recent policy initiatives in China, particularly in Zhejiang Province, offer a promising solution. Since 2018, the local government have allocated substantial budgets to settlement consolidation, and synthesized the spatial details of completed consolidation projects. By integrating these records with explanatory variables, we can now estimate consolidation potential at larger spatial scales with unprecedented precision, which represents a significant advancement in data-driven, scalable, and actionable insights into settlement consolidation.

This study conducts a spatially explicit assessment of settlement consolidation potential in the case of Zhejiang Province, China. By leveraging a machine learning method, we address challenges inherent in the consolidation potential assessment. To develop a robust PU bagging-based modeling framework, we use spatial data from 524 consolidation projects collected from local authorities as references. Model performance was optimized through hyperparameter calibration and the selection of an appropriate base classifier, validated using a 5-fold cross-validation process to minimize uncertainties. By informing location-specific consolidation practices, this study contributes to more efficient and equitable land management strategies, addressing critical challenges in food security.

2. Theoretical framework and methodology

2.1 Conceptual framework and research design

We developed a conceptual framework for settlement consolidation based on the Driving Force-Land Change (DF-C) model (Hersperger et al. 2010). This theoretical model suggests that underlying drivers are closely related to land use changes, particularly in response to external factors such as land

conflicts or policy interventions (Meyfroidt 2016; Pendrill et al. 2024). Our refined framework enhances this theoretical model by specifically focusing on the complexity of land conflicts between settlements and cropland (Fig.1).

In our model, settlement consolidation potential is determined by multiple variables. Based on a brief literature review (Supplementary Table 1), we categorized these variables into three groups: biophysical, social-economic and landscape-associated. Biophysical factors, including elevation, slope, soil properties and climate, are critical as they largely determine the quality of reclaimed cropland (Jin et al. 2017; Du et al. 2018). Social-economic factors reflect dynamics such as population decline, deteriorating building conditions, and limited transport accessibility, all of which have been shown to significantly increase consolidation potential (Xu et al. 2019; Lin et al. 2020). Landscape-associated factors, including metrics such as the aggregation index, edge density, and patch density of cropland and settlements, as well as a cropland reclamation priority index proposed by Wang et al. (2021), captures the spatial configuration of land uses. These factors are highly relevant to settlement consolidation, since isolated or fragmented land uses are more prone to convert into surrounding land uses, as observed in historical land-use changes (Zhou et al. 2017; Tao et al. 2020).

The primary principle in evaluating consolidation potential is to identify settlements with characteristics similar to those settlements that have successfully undergone consolidation. Following this principle, we employed a PU bagging approach to uncover the empirical relationship between the spatial distribution of settlements and explanatory variables. Specifically, historically consolidated settlements were designated as positive samples, while the remaining settlements were interpreted as unlabeled samples. Random subsamples selected from the unlabeled samples are integrated with positive samples to train PU bagging models, ultimately yielding a map of consolidation potential (see Section 2.3 for technical details). Moreover, this study assessed settlement consolidation potential in rural land systems of Zhejiang Province, which was delineated following an established land system classification framework (see Supplementary Materials).

Settlement consolidation offers not only a pathway to address food security challenges through

increased cropland, but also holds significant implications for environmental sustainability. In discussion section of this paper, we explored these multifaceted impacts by briefly discussing: (1) the potential crop production gains from newly reclaimed land, (2) the economic costs of consolidation, and (3) the broader environmental implications of settlement consolidation practices.

2.2 Case study

Zhejiang Province of China was selected as the study area for two primary reasons (Fig.2). On the one hand, settlement consolidation in Zhejiang Province is a pioneering initiative, serving as a flagship model in China (Liu et al. 2023; Yu et al. 2023). This province faces crucial land conflicts between crop production and settlement development, driven by economic development and population growth (Yue et al. 2022). To address these challenges, local authorities launched an innovative policy, *“Three-Year Action Plan for Comprehensive Land Consolidation and Ecological Restoration”*, in 2018 (Bryan et al. 2018). Investigating settlement consolidation potential in Zhejiang is crucial for providing scientifically grounded insights that can optimize land management practices and offer global lessons for regions facing similar conflicts. On the other hand, the availability of detailed spatial data on past consolidation projects in Zhejiang Province allows for a rigorous data-driven assessment of consolidation potential. The comprehensiveness of this dataset ensures the robustness of our research design and allows for more reliable predictions of future consolidation opportunities.

Historical settlement consolidation across Zhejiang Province showed diverse spatial patterns between 2018 and 2022 (Fig.2). During this period, 5754 hectares (ha) of settlements were consolidated, with Jiaxing (2107 ha), Huzhou (1071 ha), and Wenzhou (618 ha) leading consolidation efforts, accounting for 67% of the province's total. In contrast, Jinhua (331 ha), Lishui (182 ha), and Shaoxing (63 ha) exhibited lower levels of settlement consolidation. Notably, settlement consolidation practices were more spatially aggregated in Huzhou and Jiaxing compared with Quzhou and Taizhou.

2.3 Assessment of settlement consolidation potential

Assessing the consolidation potential involves distinguishing settlements that exhibit such potential

from those that do not. The ideal modeling process would rely on a training dataset including settlements that possess and lack consolidation potential, respectively (Bepler et al. 2019; Geng et al. 2024). However, at large spatial scales, it is challenging to effectively and efficiently determine which settlements cannot be consolidated. As a result, this modeling task represents a classic Positive-Unlabeled learning (PU learning) problem, characterized by merely having positive and unlabeled samples that are contaminated by hidden positive examples (Lee and Liu 2003; Liu et al. 2003).

Both positive and negative samples are typical modeling input for traditional supervised learning methods. While we can adapt supervised learning methods to address PU learning problem by treating all unlabeled samples as negative samples, this simplification potentially introduces biases into predictions and cannot effectively address PU learning issues (Geng et al. 2024; Gu et al. 2024). In the context of evaluating settlement consolidation potential, unlabeled settlements simultaneously include settlement possess and lack consolidation potential. Regarding unlabeled settlements with consolidation potential as negative samples can lead to class confusion and feature misrepresentation, substantially diminishing prediction accuracy (Lu and Wang 2021; Stupp et al. 2021; McDonald et al. 2021).

PU bagging is a pervasive PU learning algorithm to discriminate positive samples from unlabeled ones, and has been widely applied in land use and remote sensing studies (Lei et al. 2021; Liu et al. 2024). The underlying mechanism of PU bagging involves building multiple datasets by combining all positive samples and a down-sampled subset of unlabeled ones with replacement (Fig.1). Base classifiers (e.g., random forest or decision tree) are then trained on bootstrapped datasets and predictions are recorded for the out-of-bag (OOB) samples—those not included in the random sample. The final predictions are obtained by repeating above steps multiple times and averaging model outputs (Elkan and Noto 2008; Zhao et al. 2022). Existing studies demonstrate that PU bagging significantly reduces interference from hidden positive samples on prediction through bootstrap procedures, thereby performing better than traditional supervised learning methods (Mordelet and Vert 2014; Bepler et al. 2019; McDonald et al. 2021).

Therefore, we adopt PU bagging to assess settlement consolidation potential. We chose decision tree

and random forest as the base classifier for PU bagging. The decision tree models were set to have a maximum of 100 trees, meanwhile the subsampling and feature sampling of random forest were set as 0.5. In this study, pixels representing historically consolidated settlements were designated as positive samples for PU bagging modeling, whereas other settlements in rural areas were considered as unlabeled samples.

PU bagging training would subsequently determine hyperparameter settings, such as bagging number, down-sampling ratio and the probability threshold for classifying settlements possessing consolidation potential, through 5-fold cross-validation. Specifically, we evaluated classifier performance across varying numbers of bags (1, 10, 50 and 100), each comprising all positive samples and a random subset of unlabeled samples. The number of unlabeled samples in each bag was controlled by the down-sampling ratio, which ranged from the number of positive samples up to five times that number.

Hyperparameter settings of PU bagging would be optimized to control model complexity, reduce standard deviation and maximize the mean of the modified F1 score across folds. The modified F1 score can address evaluation challenges in PU learning, accommodating the absence of negative samples and mitigating biases in precision or F1 score calculations (Lee and Liu 2003; Norlin and Paulsrud 2017; Geng et al. 2024). The modified F1 score is calculated using positive and unlabeled samples, defined as the square of recall divided by detection prevalence (Bekker and Davis 2020; McDonald et al. 2021). This metric, which is proportional to the F1 score, ensures balanced consideration of both type 1 (False Positive) and type 2 (False Negative) errors (Jain et al. 2017; McDonald et al. 2021).

The model incorporated a wide range of explanatory variables for settlement consolidation, selected based on empirical evidence and theoretical relevance (Fig. 1 and Supplementary Table 1). These biophysical, socio-economic, and landscape-associated variables were integrated into the PU bagging model to analyze their empirical relationships to the locations of completed consolidation projects and capture the unique influence of each variable. By focusing on how these factors have historically shaped spatial distribution of consolidation practice, the model effectively informs future site

selection. To ensure robustness, we aligned the temporal scope of these variables to the year 2018, ensuring their relevance to recent consolidation activities.

2.4 Data source

Table 1 provides an overview of the datasets used in this study. Spatial details of completed consolidation projects were acquired from the local government (<https://www.zj.gov.cn/>). Land use data, derived from China land cover dataset (Yang and Huang 2021), were used to identify rural land systems in Zhejiang Province, China (Supplementary material, and Supplementary Fig.1). Additionally, land use data of 2018 was used to calculate the cropland reclamation priority index (Wang et al. 2021), and to derive landscape metrics of cropland and settlements, including aggregation index, patch density and edge density-factors that serve as explanatory variables of settlement consolidation.

Spatial layers, including building height (Wu et al. 2023), Euclidean distance to railways and roads, population density (WorldPop 2018), Euclidean distance to IUCN strict natural reserves and wilderness areas, were resampled into 30-m raster images and used as additional explanatory variables of settlement consolidation. Euclidean distance to IUCN strict natural reserves and wilderness areas was produced in 2017, while building height data was generated in 2020. Given the minimal time differences between these datasets and the study's focal year of 2018, we believe their impact on the main conclusions is marginal. To delineate possible impacts of population dynamic on settlement consolidation potential, WorldPop data were employed to calculate population change for the periods 2000-2010 and 2010-2018 (WorldPop 2018).

Terrain data, soil properties data and climate data were derived from ASTER DEMv3, SoilGrids (Poggio et al. 2021) and WorldClim (Fick and Hijmans 2017), respectively. Considering the relative stability of soil properties over short periods, we applied this data to investigate the relationship between soil properties and settlement consolidation, as minimal temporal variations are unlikely to significantly affect the analysis. Annual mean temperature, annual precipitation and solar radiation (1970-2000) were chosen to incorporate a long-term historical climate condition into the assessment of consolidation potential.

Table 1. Datasets used for our analysis. All relevant spatial datasets were resampled to 30-m raster images, which are in the Krasovsky coordinate system with a standard local Albers projection.

Variables	Time	Resolution	Data source
Settlement consolidation projects dataset	2018-2022	/	Zhejiang Province Government (https://www.zj.gov.cn/)
Land use	2018, 2022	30 m	China land cover dataset (CLCD) (https://zenodo.org/records/8176941)
Building height	2020	10 m	Wu et al. 2023 (https://zenodo.org/records/7923866)
Euclidean distance to the railway	2018	30 m	OpenStreetMap (https://www.openstreetmap.org/)
Euclidean distance to roads	2018	30 m	
Population density	2000, 2010, 2018	100 m	WorldPop (https://worldpop.org/)
Euclidean distance to IUCN strict natural reserve and wilderness area edges	2017	100 m	
Elevation	/	30 m	ASTER DEMv3 (https://www.geodata.cn)
Slope	/	30 m	
Soil organic density	2020	250 m	SoilGrids (https://www.isric.org/)
Soil pH	2020	250 m	
Soil nitrogen	2020	250 m	
Annual mean temperature	1970-2000	1000 m	WorldClim version 2.1 (https://www.worldclim.org/)
Annual precipitation			
Solar radiation			

3. Results

3.1 Spatial distribution of settlement consolidation potential

Our assessment found 29,784 ha of settlements possessing consolidation potential in Zhejiang Province (Fig. 3a). A large proportion of this potential is concentrated in Jiaxing (12,961 ha), Huzhou (8987 ha), and Wenzhou (3340 ha), representing 44%, 30%, and 11% of the total, respectively. Furthermore, these areas are more spatially contiguous in Jiaxing compared with Wenzhou.

The average settlement area among rural land systems is 5.44 ha, while the average potential consolidation area is 0.67 ha (Fig.3b). Regions like Lishui, Shaoxing, Zhoushan and Ningbo show limited consolidation potential, with less than 5% of land systems exceeding 1 ha of potential consolidation area. In contrast, rural systems in Quzhou, Huzhou and Jiaxing exhibit significant consolidation opportunities, with 2%, 14%, and 26% of land systems containing more than 5 ha of settlements with consolidation potential, respectively.

Rural land systems with smaller settlement size exhibits high consolidation potential (Fig.4). Across Zhejiang Province, 71% of rural land systems exhibit limited consolidation potential, with less than 1% of the total consolidation potential. Nonetheless, over 18% rural land systems still offer significant consolidation opportunities, with over half of their settlements suitable for consolidation. A clear trend is observed within the 0–25 ha settlement area range: as settlement size increases, the mean proportion of settlement with consolidation potential decreases. For instance, settlements between 0–5 ha exhibit a mean consolidation proportion of 24%, which drops sharply to 9% for those in the 15–20 ha range.

Our results present a distinct pattern of settlement consolidation potential across plains, hills, and mountains (Fig.4). Plains account for almost 80% of the potential area for settlement consolidation in rural areas, followed by mountains (17%) and hills (3%). For land systems with over 75% of settlements exhibiting consolidation potential, rural land systems in plains have the highest average settlement size, at 6.18 ha, which is substantially larger than the 0.78 ha in mountains and 0.54 ha in hills. Despite these contrasts, a common trend emerges across all geomorphic divisions: rural land

systems with high consolidation potential—where over 90% of settlements are identified as potential consolidation areas—are predominantly associated with settlement sizes smaller than 2 ha.

The settlement consolidation potential is influenced by various biophysical, socio-economic, and landscape-associated factors, each exhibiting different levels of influences (Fig. 5). Specifically, factors such as distance to conservation areas, annual mean temperature, and annual precipitation are more influential than others. In contrast, most landscape-associated factors, except for the aggregation index of settlement, have relatively limited influences. Solar radiation was excluded in final prediction due to low feature importance.

3.2 Model performance and uncertainties

We assess the settlement consolidation potential at a 30 meters resolution using ensemble models through 5-fold cross-validation (Fig. 6). Our results indicate that PU bagging model with decision tree as the base classifier generally achieves higher recall scores compared to random forest models across most parameter settings. However, random forest models consistently exhibit higher detection prevalence scores. Considering the modified F1 scores, choice of decision tree models as base classifier outperforms random forest models, demonstrating superior performance in identifying settlements possessing consolidation potential. Therefore, the decision tree was selected as base classifier for the PU bagging model. Furthermore, the decision tree-based PU bagging models demonstrates relatively stable modified F1 scores across different down-sampling ratios, with a slight improvement as the numbers of bagging iterations increased. Accordingly, the optimized down-sampling ratio and bagging numbers are determined to be 1 and 100, respectively.

To reduce uncertainties in classification of settlement with consolidation potential, we optimized probability thresholds within the PU bagging model, focusing on maximize the modified F1 score. As the model outputs probabilities indicating the likelihood of consolidation, selecting an appropriate threshold is critical to determining which settlement cells are classified as having consolidation potential. Through a comprehensive test of thresholds ranging from 0.01 to 1 during ensemble modeling, we identified an optimal threshold of 0.52 for the final prediction model. This threshold

yielded a modified F1 score of 4.05, a recall of 0.75, and a detection prevalence of 0.14 across 5-fold cross-validation (Fig. 7a).

We estimate the coefficient of variation (CV) of settlement consolidation probability in repeated 50 model runs to quantify the uncertainties of PU bagging model. The CV of consolidation probability exhibit an inverse-U pattern along with the increase of consolidation probability (Fig. 7b and Supplementary Fig. 5). The uncertainty in Zhejiang Province is specifically found most in low consolidation probability (around 0.3-0.5), whereas the CV would meet a rapid decline as consolidation probability goes beyond around 0.5 (Fig. 7b and Supplementary Fig. 5). Besides, the inverse-U dynamic pattern of CV is observed across three geomorphic divisions. Although the highest value of CV in plains is larger than Hills and Mountains, modeling uncertainties in plains would meet more significant reduction in consolidation probability around 0.4 to 0.7.

4. Discussion and conclusion

4.1 Settlement consolidation as a solution for food insecurity and beyond

Settlement consolidation offers a promising solution to offset cropland loss and boost food production, particularly in the context of global food security. Our results explicitly illustrate the spatial heterogeneity in settlement consolidation potential. Here, we identify 29,784 ha potential areas for settlement consolidation in Zhejiang Province, predominantly in Jiaxing (12,961 ha) and Huzhou (8987 ha). From the perspective of geomorphic divisions, plains stand out, accounting for almost 80% potential area for settlement consolidation (Fig. 4), followed by mountains (17%) and hills (3%). Moreover, implementing consolidation in these areas can increase local food production by 0.91–0.97% (53,306–55,558 tons) compared to the 2022 levels (Supplementary Fig.3). While this growth may seem modest, it holds strategic importance in offsetting ongoing declines in crop production and improving food security.

In addition to promoting food production, consolidation potential map can explicitly support the integration of settlement consolidation into multi-objective planning (Strassburg et al. 2020; Bateman et al. 2024). Food security, biodiversity conservation, and climate change mitigation are all integral

elements of the Sustainable Development Goals (Salerno et al. 2024). Therefore, stakeholders need decision-making support to achieve these multiple objectives and reconcile potential trade-offs between policy responses (Bai et al. 2018). For instance, cropland expansion or intensification are widely employed to meet challenges on global food system from climate change, population dynamic and dietary change (Willett et al. 2019), while these measures are proven to potentially intensify global land competition between cropland and natural habitat, causing habitat loss and degradation (van Vliet 2019; Williams et al. 2020). Previous studies indicate that settlement consolidation can contribute to biodiversity conservation and carbon emissions reduction by optimizing rural landscapes (Yue et al. 2023; Francini et al. 2024). Our study further demonstrates that settlement consolidation can serve as an alternative pathway to address significant increasing food demand without compromising conservation objectives by mitigating the trade-offs between food security and ecosystem conservation.

4.2 Towards effective evaluation of settlement consolidation potential

This study introduces a scalable and robust tool for evaluating settlement consolidation potential at large spatial scales. Traditional methods, such as interviews or multi-criteria evaluation models, are often confined to smaller spatial scales or lack of robustness (Zhou et al. 2017; Pašakarnis et al. 2021). Similarly, employing conventional machine learning method would introduce significant biases into consolidation potential evaluation. To address these challenges, we leverage the PU bagging approach specifically designed for scenarios where only positive and unlabeled samples are available. This method employs positive samples and random subsamples from unlabeled data to analyze the empirical relationship between training samples and underlying consolidation drivers, enabling the development of a generic model capable of identifying potential consolidation areas within unlabeled settlement datasets. To further enhance the model's accuracy and reliability, our modeling framework incorporates ensemble modeling, hyperparameter tuning, and an optimization algorithm for consolidation probability thresholds. These methodologies significantly improve prediction accuracy by maximizing recall and the modified F1 score while minimizing the coefficient of variation, thereby reducing modeling uncertainty. By providing precise spatial assessments, our innovative evaluation

framework delivers valuable insights to policymakers and supports informed decision-making, paving the way for more location-specific and effective land use strategies.

Stakeholder decisions are pivotal in land use changes (Malek and Verburg 2020). In the context of settlement consolidation, key actors—such as local governments, villagers, and other stakeholders—make decisions based on the aforementioned explanatory variables of settlement consolidation and decision preferences (Liu et al. 2020; Gao et al. 2022b). Traditionally, understanding these preferences required extensive in-person consultations, which are time-consuming and costly (Tian et al. 2018; Zhao et al. 2024b). In contrast, machine learning algorithms offer a more effective and reliable approach by analyzing the relationship between implemented consolidation projects and explanatory variables (Mordelet and Vert 2014; Stupp et al. 2021). By learning from the characteristics of areas where consolidation has already occurred, these models can identify new areas with similar features that align with these established patterns. Since the location of implemented projects reflects stakeholder decisions preference, we reasonably assume that the areas identified by the machine learning model would also align with these stakeholders' decisions preference. Additionally, the model's high recall score validates this assumption, indicating its robustness in identifying consolidation sites shaped by stakeholder decisions (Fig.4 and Supplementary Fig.3). These results emphasize the model's effectiveness in reflecting decision-making preference, ensuring alignment between consolidation location selection with local preferences.

4.3 Policy implications

Financial investment is critically relevant to settlement consolidation. To assess the economic feasibility of such an effort, we further evaluated required financial investment by using empirical cost data from previous consolidation projects in Zhejiang Province. On average, settlement consolidation in Zhejiang Province costs 0.48 million USD per ha, encompassing financial compensation for displaced homeowners, infrastructure development for agriculture (e.g., roads and water supply), project planning, and design expenses (Supplementary Fig.4). Moreover, consolidating all potential areas in Zhejiang Province would require a total investment of approximately 18.06 billion

USD. The investment is expected to be unevenly distributed across space (Supplementary Fig.4), which further demonstrates the necessity of spatial assessment of consolidation potential to carefully and strategically implement settlement consolidation to maximize benefits while minimizing cost.

Currently, substantial subsidies for consolidation projects are provided by local authorities, primarily funded through local tax revenues. However, to enhance economic incentives and ensure long-term financial sustainability, the Zhejiang provincial government has introduced a market-driven mechanism. Under the policy outlined in *“Zhejiang Province’s Opinions on Implementing Comprehensive Land Consolidation and Ecological Restoration Projects”* (ZPG 2018), land quotas generated from settlement consolidation—analogueous to transferable development rights—can now be traded in the market (Chen et al. 2023). This innovative approach not only addresses funding gaps but also fosters broader participation in settlement consolidation.

Settlement consolidation goes beyond demolishing old houses and reclaiming cropland, it has profound implications for local well-beings, livelihood, and cultural heritage of local communities. Housing is a critical factor in economic and social stability for these families, significantly influencing their willingness to participate in consolidation practices (Liu et al. 2020; Gao et al. 2023). Thus, regions with high consolidation potential are supposed to implement comprehensive welfare programs to compensate residents who lose their residences and ensure the maintenance of desired livelihoods after consolidation. Other than economic dimension, some settlements hold significant cultural and historical value. Many old buildings in rural areas are important historical assets that deserve deliberate protection, to meet cultural preservation goals and sustain historical identity of these communities. Interviews reveal that for some residents, settlement serve as emotional anchors, sites of ancestral worships, and centers of family connection (Fang et al. 2016). Consequently, settlement consolidation should prioritize minimizing land use conflicts and creating synergies between settlement development, local well-being, and cultural protection.

References

- Bai Y, Wong CP, Jiang B, et al (2018) Developing China's Ecological Redline Policy using ecosystem services assessments for land use planning. *Nat Commun* 9:1–13. <https://doi.org/10.1038/s41467-018-05306-1>
- Bateman IJ, Binner A, Addicott ET, et al (2024) How to make land use policy decisions: Integrating science and economics to deliver connected climate, biodiversity, and food objectives. *Proc Natl Acad Sci* 121:e2407961121. <https://doi.org/10.1073/pnas.2407961121>
- Bekker J, Davis J (2020) Learning from positive and unlabeled data: A survey. *Mach Learn* 109:719–760
- Bepler T, Morin A, Rapp M, et al (2019) Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat Methods* 16:1153–1160. <https://doi.org/10.1038/s41592-019-0575-8>
- Bryan BA, Gao L, Ye Y, et al (2018) China's response to a national land-system sustainability emergency. *Nature* 559:193–204. <https://doi.org/10.1038/s41586-018-0280-2>
- Cay T, Uyan M (2013) Evaluation of reallocation criteria in land consolidation studies using the Analytic Hierarchy Process (AHP). *Land use policy* 30:541–548. <https://doi.org/10.1016/j.landusepol.2012.04.023>
- Chen M, Yao S, Hu C, Jin S (2023) Transfer or retain land development right: The role of China's IDB programme in supporting inclusive urbanisation. *Urban Stud* 60:2651–2668. <https://doi.org/10.1177/00420980231155026>
- Christopher, Reitsma F, Baiocchi G, et al (2017) Future urban land expansion and implications for global croplands. *Proc Natl Acad Sci* 114:201606036. <https://doi.org/10.1073/pnas.1606036114>
- Colombo S, Perujo-Villanueva M (2019) A practical method for the ex-ante evaluation of land consolidation initiatives: Fully connected parcels with the same value. *Land use policy* 81:463–471. <https://doi.org/10.1016/j.landusepol.2018.11.018>
- Daskalova GN, Kamp J (2023) Abandoning land transforms biodiversity. *Science* (80-) 380:581–583. <https://doi.org/10.1126/science.adf1099>
- Dong Y, Cheng P, Kong X (2022) Spatially explicit restructuring of rural settlements: A dual-scale coupling approach. *J Rural Stud* 94:239–249. <https://doi.org/10.1016/j.jrurstud.2022.06.011>
- Du X, Zhang X, Jin X (2018) Assessing the effectiveness of land consolidation for improving agricultural productivity in China. *Land use policy* 70:360–367. <https://doi.org/10.1016/j.landusepol.2017.10.051>
- Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp 213–220
- Fang Y, Shi K, Niu C (2016) A comparison of the means and ends of rural construction land consolidation: Case studies of villagers' attitudes and behaviours in Changchun City, Jilin province, China. *J Rural Stud* 47:459–473. <https://doi.org/10.1016/j.jrurstud.2016.04.007>
- Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 37:4302–4315. <https://doi.org/10.1002/joc.5086>

- Francini S, Chirici G, Chiesi L, et al (2024) Global spatial assessment of potential for new peri-urban forests to combat climate change. *Nat Cities* 1:286–294. <https://doi.org/10.1038/s44284-024-00049-1>
- Gao J, Cai Y, Liu Y, et al (2022a) Understanding the underutilization of rural housing land in China: A multi-level modeling approach. *J Rural Stud* 89:73–81. <https://doi.org/10.1016/j.jrurstud.2021.11.020>
- Gao J, Cai Y, Wen Q, et al (2023) Future matters: Unpacking villagers’ willingness to withdraw from rural homesteads in China. *Appl Geogr* 158:103049. <https://doi.org/10.1016/j.apgeog.2023.103049>
- Gao J, Song G, Liu S (2022b) Factors influencing farmers’ willingness and behavior choices to withdraw from rural homesteads in China. *Growth Change* 53:112–131. <https://doi.org/10.1111/grow.12574>
- Geng J, Ding Y, Xie W, et al (2024) An ensemble machine learning model to uncover potential sites of hazardous waste illegal dumping based on limited supervision experience. *Fundam Res* 4:972–978. <https://doi.org/https://doi.org/10.1016/j.fmre.2023.06.010>
- Gu T, Duan P, Wang M, et al (2024) Effects of non-landslide sampling strategies on machine learning models in landslide susceptibility mapping. *Sci Rep* 14:7201. <https://doi.org/10.1038/s41598-024-57964-5>
- Hersperger AM, Gennaio M, Verburg PH (2010) Linking Land Change with Driving Forces and Actors: Four Conceptual Models. *Ecol Soc* 15:1
- Jain S, White M, Radivojac P (2017) Recovering true classifier performance in positive-unlabeled learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*
- Janus J, Markuszewska I (2019) Forty years later: Assessment of the long-lasting effectiveness of land consolidation projects. *Land use policy* 83:22–31. <https://doi.org/10.1016/j.landusepol.2019.01.024>
- Jin X, Shao Y, Zhang Z, et al (2017) The evaluation of land consolidation policy in improving agricultural productivity in China. *Sci Rep* 7:2792. <https://doi.org/10.1038/s41598-017-03026-y>
- Jürgenson E (2016) Land reform, land fragmentation and perspectives for future land consolidation in Estonia. *Land use policy* 57:34–43. <https://doi.org/10.1016/j.landusepol.2016.04.030>
- Lambin EF, Kim H, Leape J, Lee K (2020) Scaling up Solutions for a Sustainability Transition. *One Earth* 3:89–96. <https://doi.org/https://doi.org/10.1016/j.oneear.2020.06.010>
- Lee WS, Liu B (2003) Learning with positive and unlabeled examples using weighted logistic regression. In: *ICML*. pp 448–455
- Lei L, Wang X, Zhong Y, et al (2021) DOCC: Deep one-class crop classification via positive and unlabeled learning for multi-modal satellite imagery. *Int J Appl Earth Obs Geoinf* 105:102598. <https://doi.org/10.1016/j.jag.2021.102598>
- Lin S-H, Liu K-M, Hsieh J-C, et al (2020) A new hybrid modified MADM model for the potential evaluation of a comprehensive land consolidation project (LCP) toward achieving sustainable development. *J Environ Plan Manag* 63:1585–1615. <https://doi.org/10.1080/09640568.2019.1673152>
- Liu B, Dai Y, Li X, et al (2003) Building text classifiers using positive and unlabeled examples. In: *Third IEEE international conference on data mining*. IEEE, Melbourne, FL, USA, pp 179–186

- Liu R, Yu C, Jiang J, et al (2020) Farmer differentiation, generational differences and farmers' behaviors to withdraw from rural homesteads: Evidence from chengdu, China. *Habitat Int* 103:102231. <https://doi.org/10.1016/j.habitatint.2020.102231>
- Liu Y, Dai L, Long H (2023) Theories and practices of comprehensive land consolidation in promoting multifunctional land use. *Habitat Int* 142:102964. <https://doi.org/10.1016/j.habitatint.2023.102964>
- Liu Y, Fang F, Li Y (2014) Key issues of land use in China and implications for policy making. *Land use policy* 40:6–12. <https://doi.org/10.1016/j.landusepol.2013.03.013>
- Liu Z, Zhao H, Wang X, et al (2024) PU-KBS: A Robust Positive and Unlabeled Learning Framework with Key Band Selection for One-Class Hyperspectral Image Classification. *IEEE Trans Geosci Remote Sens* 32:5708–5724. <https://doi.org/10.1109/TGRS.2024.3397989>
- Lu Y, Wang L (2021) How to automate timely large-scale mangrove mapping with remote sensing. *Remote Sens Environ* 264:112584. <https://doi.org/10.1016/j.rse.2021.112584>
- Malek Ž, Verburg PH (2020) Mapping global patterns of land use decision-making. *Glob Environ Chang* 65:102170. <https://doi.org/10.1016/j.gloenvcha.2020.102170>
- McDonald GG, Costello C, Bone J, et al (2021) Satellites can reveal global extent of forced labor in the world's fishing fleet. *Proc Natl Acad Sci* 118:e2016238117. <https://doi.org/10.1073/pnas.2016238117>
- Meyfroidt P (2016) Approaches and terminology for causal analysis in land systems science. *J Land Use Sci* 11:501–522. <https://doi.org/10.1080/1747423X.2015.1117530>
- Mordelet F, Vert JP (2014) A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit Lett* 37:201–209. <https://doi.org/10.1016/j.patrec.2013.06.010>
- NLPC (2017) National Land Planning of China (2016-2030). In: State Counc. People's Repub. China. https://www.gov.cn/zhengce/content/2017-02/04/content_5165309.htm
- Norlin P, Paulsrud V (2017) Identifying New Customers Using Machine Learning: A case study on B2B-sales in the Swedish IT-consulting sector. KTH Royal Institute of Technology
- Pan H, Wu Y, Choguill C (2023) Optimizing the rural comprehensive land consolidation in China based on the multiple roles of the rural collective organization. *Habitat Int* 132:102743. <https://doi.org/10.1016/j.habitatint.2023.102743>
- Pašakarnis G, Maliene V, Dixon-Gough R, Malys N (2021) Decision support framework to rank and prioritise the potential land areas for comprehensive land consolidation. *Land use policy* 100:104908. <https://doi.org/10.1016/j.landusepol.2020.104908>
- Pendrill F, Gardner TA, Meyfroidt P, et al (2024) Disentangling the numbers behind agriculture-driven tropical deforestation. *Science* (80-) 377:eabm9267. <https://doi.org/10.1126/science.abm9267>
- Poggio L, De Sousa LM, Batjes NH, et al (2021) SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* 7:217–240. <https://doi.org/10.5194/soil-7-217-2021>
- Salerno J, Gaughan AE, Warrier R, et al (2024) Rural migration under climate and land systems change. *Nat Sustain* 1–10.

<https://doi.org/10.1038/s41893-024-01396-6>

- Strassburg BBN, Iribarrem A, Beyer HL, et al (2020) Global priority areas for ecosystem restoration. *Nature* 586:724–729. <https://doi.org/10.1038/s41586-020-2784-9>
- Stupp D, Sharon E, Bloch I, et al (2021) Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat Commun* 12:6454. <https://doi.org/10.1038/s41467-021-26792-w>
- Tao Z, Guanghui J, Guangyong L, et al (2020) Neglected idle rural residential land (IRRL) in metropolitan suburbs: Spatial differentiation and influencing factors. *J Rural Stud* 78:163–175. <https://doi.org/10.1016/j.jrurstud.2020.06.020>
- Tao Z, Guanghui J, Wenqiu M, et al (2024) A framework for identifying the distribution of revitalization potential of idle rural residential land under rural revitalization. *Land use policy* 136:106977. <https://doi.org/10.1016/j.landusepol.2023.106977>
- Tian Y, Liu Y, Kong X (2018) Restructuring rural settlements based on mutualism at a patch scale: A case study of Huangpi District, central China. *Appl Geogr* 92:74–84. <https://doi.org/10.1016/j.apgeog.2018.01.015>
- van Vliet J (2019) Direct and indirect loss of natural area from urban expansion. *Nat Sustain* 2:755–763. <https://doi.org/10.1038/s41893-019-0340-0>
- Wang S, Bai X, Zhang X, et al (2021) Urbanization can benefit agricultural production with large-scale farming in China. *Nat Food* 2:183–191. <https://doi.org/10.1038/s43016-021-00228-6>
- Willett W, Rockström J, Loken B, et al (2019) Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. *Lancet* 393:447–492. [https://doi.org/10.1016/S0140-6736\(18\)31788-4](https://doi.org/10.1016/S0140-6736(18)31788-4)
- Williams DR, Clark M, Buchanan GM, et al (2020) Proactive conservation to prevent habitat losses to agricultural expansion. *Nat Sustain* 4:314–322. <https://doi.org/10.1038/s41893-020-00656-5>
- WorldPop (2018) Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). <https://doi.org/10.5258/SOTON/WP00645>
- Wu W-B, Ma J, Banzhaf E, et al (2023) A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sens Environ* 291:113578. <https://doi.org/10.1016/j.rse.2023.113578>
- Xu F, Ho HC, Chi G, Wang Z (2019) Abandoned rural residential land: Using machine learning techniques to identify rural residential land vulnerable to be abandoned in mountainous areas. *Habitat Int* 84:43–56. <https://doi.org/10.1016/j.habitatint.2018.12.006>
- Yang J, Huang X (2021) The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth Syst Sci Data Discuss* 2021:1–29. <https://doi.org/10.5194/essd-13-3907-2021>
- Yang R, Chen S, Ye Y (2024) Toward potential area identification for land consolidation and ecological restoration: an integrated framework via land use optimization. *Environ Dev Sustain* 26:3127–3146. <https://doi.org/10.1007/s10668-022-02767-9>
- Yu W, Yang J, Wu F, et al (2023) Realistic characteristics and driving mechanisms of pseudo-human settlements in Chinese

cities. *Humanit Soc Sci Commun* 10:50. <https://doi.org/10.1057/s41599-023-01551-y>

- Yue W, Feng B, Zhou Q, et al (2022) An assessment of the Ecological Conservation Redline: unlocking priority areas for conservation. *J Environ Plan Manag* 67:1034–1052. <https://doi.org/10.1080/09640568.2022.2145939>
- Yue W, Zhou Q, Li M, Vliet J Van (2023) Relocating built-up land for biodiversity conservation in an uncertain future. *J Environ Manage* 345:118706. <https://doi.org/10.1016/j.jenvman.2023.118706>
- Zang Y, Yang Y, Liu Y (2021) Toward serving land consolidation on the table of sustainability: An overview of the research landscape and future directions. *Land use policy* 109:105696. <https://doi.org/10.1016/j.landusepol.2021.105696>
- Zhao L, Liu G, Lu Z, et al (2024a) A new framework for delineating farmland consolidation priority areas for promoting agricultural mechanization in hilly and mountainous areas. *Comput Electron Agric* 218:108681. <https://doi.org/10.1016/j.compag.2024.108681>
- Zhao X, Cai B, He J, Kong X (2024b) Identifying potential rural residential areas for land consolidation using a data driven agent-based model. *Land use policy* 145:107260. <https://doi.org/10.1016/j.landusepol.2024.107260>
- Zhao Y, Xu Q, Jiang Y, et al (2022) Dist-pu: Positive-unlabeled learning from a label distribution perspective. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp 14461–14470
- Zhou J, Qin X, Liu L, Hu Y (2017) A potential evaluation model for land consolidation in fragmental regions. *Ecol Indic* 74:230–240. <https://doi.org/10.1016/j.ecolind.2016.09.008>
- ZPG (2018) Zhejiang Province's Opinions on Implementing Comprehensive Land Consolidation and Ecological Restoration Projects. In: Zhejiang Prov. Gov. http://sthjt.zj.gov.cn/art/2020/6/19/art_1229129343_47620138.html

Figure legends

Figure 1. Conceptual illustration of settlement consolidation

Figure 2. Study area. The blue polygons are implemented settlement consolidation projects (2018-2022). Credit: World Ocean Base by ESRI.

Figure 3. Consolidation potential of settlements. (a) represents the potential area for settlement consolidation; (b) displays the summarized potential areas for settlement consolidation at a 30-meter resolution, aggregated to each land system grid ($0.9 \times 0.9 \text{ km}^2$). Credit: World Ocean Base by ESRI.

Figure 4. Relative frequency distribution of settlement pixels as the joint occurrence of settlement sizes and the proportion of settlements with consolidation potential in rural land systems, shown for Zhejiang Province and its three geomorphic divisions: Plains, Hills, and Mountains. The delineation of geomorphic divisions is shown in Supplementary Figure 2.

Figure 5. Mean feature importance of explanatory variables for settlement consolidation across bags. Feature importance results cannot offer any information about the directionality of each variable's ability to accurately identify settlements with consolidation potential.

Figure 6. Model performance in recall, detection prevalence, and modified F1 score of cross-folder validation. The panels are organized horizontally by the number of bags employed, with performance metrics displayed vertically. The x-axis represents increasing down-sampling ratios, and the y-axis indicates the mean value of each performance metric across 5-fold cross-validation, with error bars

representing ± 1 standard deviation. Each model variation is evaluated using the optimized threshold that maximizes the modified F1 score.

Figure 7. Model uncertainty assessments. (a) represents numbers of settlement pixels that final model classifies as positive (settlement with consolidation potential), broken apart by initial training data set label (positive or unlabeled), where the positive label means this settlement has been consolidated during 2018-2022. (b) illustrates the coefficient of variation (CV) of settlement consolidation probability as a function of its values. The CV is calculated as the ratio of the standard deviation to the mean settlement consolidation probability for each cell across 50 runs. A wide range suggests significant variation among models, while a narrow range reflects greater agreement, implying a more reliable estimate. The shaded areas indicate interquartile ranges (25%–75%), and the lines represent the ensemble median values. The optimized cutoff threshold for determining consolidation potential is presented as dotted vertical line.

Data availability

The raw data sources are detailed in Table 1. The datasets generated during the current study are available in the Figshare repository, <https://doi.org/10.6084/m9.figshare.28711544>. For interactive exploration, the data can also be freely visualized at <https://www.landbigdata.info/SCP/>.

Acknowledgement

We would like to acknowledge Professor Yong Liu for the valuable discussion. This study was supported by the National Natural Science Foundation of China (No. 42471286) and the Fundamental Research Funds for the Central Universities. Q. Zhou was supported by China Scholarship Council (No.202306320269).

Author contributions

Conceptualization: QZ, WY and ML; data curation: QZ, WY and LZ; formal analysis: QZ; writing—original draft: QZ; software: QZ; writing—review and editing: QZ, WY, ML and HH. All authors have read and agreed to the published version of the manuscript.

Ethics declarations

Competing interests. The authors declare no competing interests.

Ethical approval. This article does not contain any studies with human participants performed by any of the authors.

Informed consent. This article does not contain any studies with human participants performed by any of the authors.