Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Parcel-Level PPO Framework for County-Scale Soybean–Corn Strip Intercropping Expansion under Subsidy and Site-Selection Compliance

Niantang Liu

liuniantang@caas.cn

Chinese Academy of Agricultural Sciences

Guijun Yang

Beijing Academy of Agriculture and Forestry Sciences

Xiaodong Yang

Beijing Academy of Agriculture and Forestry Sciences

Si-Bo Duan

Chinese Academy of Agricultural Sciences

Liang Sun

Chinese Academy of Agricultural Sciences

Zixuan Liu

Chinese Academy of Agricultural Sciences

Research Article

Keywords: strip intercropping, cropping structure adjustment, reinforcement learning, Proximal Policy Optimization, parcel-level decision support, Huang–Huai–Hai Plain, agricultural subsidy, site-selection compliance

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9950327/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Soybean–corn strip intercropping is a national strategy for easing China's protein-crop deficit while protecting wheat. Turning a county pledge (Huojia County's 3,000-to-10,000 mu expansion; 1 mu ≈ 0.0667 ha) into a parcel-level plan couples a band-shaped area target, wheat protection, threshold-based site-selection compliance, and a non-linear subsidy-augmented economic objective. We propose a parcel-level Proximal Policy Optimization (PPO) framework that keeps every cultivated non-strip parcel, including every rule violator, in the decision pool, with compliance entering only through the reward via an auditable rate ρ and an adjustable acceptance floor ρ* (default 0.85). Because per-mu conversion gains are largest for rule-violating parcels, compliance must be learned against the economic gradient, and a counterfactual (difference-reward) advantage estimator repairs the credit-assignment failure of episode-level PPO over 8,833 simultaneous decisions. On Huojia County (17,833 parcels), the learned policy designates 31 parcels (8,828 mu) within the 7,000–10,000 mu band, preserves the wheat floor, and raises gross margin by 3.79 million yuan (+ 3.4%) over the 2022 baseline; it audits at ρ = 0.871 versus 0.000 for a profit-greedy heuristic on the same pool, separating compliant from violating parcels almost perfectly (AUC ≈ 1.0) across three seeds. Compliance tracks the dial: ρ* = 0.95 yields 0.957 and a zero-tolerance fine yields exactly 1.000. A neighbourhood-conditioned contiguity variant reaches up to 76% spatial coherence at matched area, versus 6% for a size-sorted heuristic that moreover overshoots the band. We claim no economic superiority: a correctly targeted size-sorted heuristic ties the learned plan, and under zero tolerance the policy reproduces that heuristic's identical 8,601 mu / 27-parcel plan; the contribution is that compliance, the tunable floor, and the contiguity coupling emerge jointly from one reward formulation rather than a hand-chosen sort key. The framework outputs GIS-ready designation shapefiles; a subsidy-response sweep and a network-architecture comparison remain future work.

1. Introduction

China's grain policy has entered a new phase in which the long-standing emphasis on cereal self-sufficiency is being supplemented by an explicit push for protein-crop expansion [1, 2]. The expansion is driven by structural imbalances: a domestic feed-grain deficit met largely through soybean imports, livestock-product demand growing faster than the corresponding feed-grain supply, and external-circulation risks in international trade. At the same time, food-security guidelines preclude the simple substitution of soybean for wheat or rice in regions where those crops anchor regional supply chains [3–5]. The mismatch between current cropping patterns and underlying resource endowments has been documented at multiple scales, from county-level cropping-structure adjustments [3, 6] through climate-driven boundary shifts [7, 8] to global cropping-intensity gaps [5], and it motivates targeted, site-specific policy instruments rather than uniform regional directives.

This spatial mismatch is well established empirically. The structure of Chinese crop production has shifted dramatically since the 1980s: Liu et al. [3] document a decisive 2002 inflection at which the rice-, wheat-, and corn-dominated cropping pattern transitioned into a more diverse mix incorporating fruit and vegetable types, with 47% of rice-, 61% of wheat-, and 29.6% of corn-producing counties showing

statistically significant declines while soybean expanded significantly in 50.3% of counties. Tang et al. [6] survey the remote-sensing-based monitoring of these spatial patterns, and Liu et al. [7] document the climate-driven northern-limit shift of maize planting in Northeast China; comparable diversity dynamics have been characterized for the United States [9, 10] and through the global gridded cropland studies of Leff et al. [11] and the rice-cropland-trajectory work of Zhang et al. [12]. Building on the Global Agro-Ecological Zones framework [13], Li [1] constructs potential-yield development coefficients and yield-efficiency advantage indices, finding that roughly 26.42% of national soybean and 34.74% of national corn were planted in inappropriate regions during the study period, while Song and Zhang [14] apply an agent-based land-allocation model with multi-criterion suitability evaluation to derive optimized rice–maize–soybean layouts at the county level in Heilongjiang's main grain belt. These works collectively establish that current Chinese cropping structures are spatially misaligned with resource endowments [4, 5], yet they stop short of providing the parcel-level designation tool needed for operational pilot implementation.

Soybean–corn strip intercropping has emerged as the policy instrument of choice because it allows soybean acreage to increase without proportionally displacing the corn area on which feed-grain volume depends. The configuration promoted in Henan alternates 4–6 rows of soybean with 2–4 rows of corn (the 4:2 and 6:4 patterns of the provincial technical guidance), exploits spatial niche complementarity to deliver yield-per-area performance comparable to monoculture corn [15], and is compatible with existing mechanization. The challenge is operational: converting a national or provincial policy target into a parcel-level designation plan at the county level. A typical Huang–Huai–Hai county contains tens of thousands of cultivated parcels with heterogeneous size, prior-crop history, and locational suitability, and existing decision-support tools, whether manual review by extension officers, ad hoc spreadsheets, or county-level linear-programming models [16–23], struggle on three fronts. First, the area target is naturally band-shaped (e.g., 7,000 to 10,000 mu of new designation in our case study, encoding a pledge to expand from 3,000 to 10,000 mu) rather than a fixed number, which is awkward to encode in linear-programming objectives. Second, site-selection compliance is conditional and threshold-based (continuous-soybean if soybean ratio ≥ 0.3, small-parcel if area < 20 mu, isolated-location if its 2022 mode is not in the permissible set {wheat–corn, strip}), translating into a combinatorial constraint set with weak linear relaxations. Third, the subsidy-augmented economic surface is non-linear, because the strip-intercropping per-mu profit far exceeds the alternative wheat–corn double-cropping profit only when the designated area falls within the subsidy-eligible band.

These difficulties expose the limits of the established crop-planning toolbox. Two families of population-based metaheuristics dominate the multi-objective literature: particle swarm optimization (PSO), introduced by Kennedy and Eberhart [24] and extended to multi-objective problems by Coello et al. [25], with recent regional applications such as the multi-objective PSO of Begam et al. [26]; and the elitist non-dominated sorting genetic algorithm NSGA-II [27], applied to crop planning by Sarker and Ray [28]. Both produce dense Pareto frontiers but encode threshold-style compliance constraints through fitness penalties and operate on a flat binary vector that does not exploit per-parcel heterogeneity. The mathematical-programming tradition is similarly aggregate: Chen et al. [29] develop the EESO-CP-CC

model, which couples a multi-objective programming framework with a non-point-source pollution input-output model derived from Johnes [30] and the foundational stream-purification work of Streeter and Phelps [31], applied to the Jinxi Irrigation District in Heilongjiang, but operates at the irrigation-district scale with continuous decision variables rather than combinatorial parcel-level designation. Jain et al. [16] review the broader landscape of linear programming, fuzzy goal programming, and MOTAD-style approaches, reaching back to the foundational sector-planning typology of Norton and Schiefer [20], and complementary formulations span virtual-water and energy exchanges [32], fuzzy multi-objective linear-fractional programming [33], water-footprint accounting [34], groundwater-aware planning [18, 22], fuzzy programming [19], fuzzy goal programming [23], groundwater-sustainable plans [17], and further national crop-mix, fuzzy multi-objective, and multiple-cropping simulation studies [21, 35–37]. This body of work establishes the analytical basis but generally operates at regional or irrigation-district scale on flat encodings and cannot express per-parcel threshold compliance.

A further strand of literature supplies the policy rationale, the data infrastructure, and the algorithmic paradigm on which a parcel-level solver can build. On the policy-response side, Zheng et al. [38] use endogenous switching regression on 2015–2016 Heilongjiang panel data to study the heterogeneous response of corn producers to the 2016 corn price-support reform, finding that the rise in soybean area occurred more through area expansion than through corn–soybean substitution, consistent with continuous-corn rotation prevailing under farmer profit incentives and paralleling the US rotational evidence of Hendricks et al. [39]; crucially, crop structure for large-scale producers proves more price-responsive than for medium- and small-scale producers, reinforcing the operational necessity of parcel-level rather than aggregate designation, with the underlying selectivity-model apparatus drawn from Goetz [40], Key et al. [41], Lokshin and Sajaia [42], and Wooldridge [43]. On the data side, the parcel records that anchor an operational framework are increasingly available through high-resolution remote sensing: Yin et al. [44] propose an automatic spectro-temporal feature selection method on Sentinel-2 time-series imagery to map rice, corn, and soybean in the Sanjiang Plain at 93.94% overall accuracy, building on the broader Chinese crop-pattern mapping literature surveyed by Hu et al. [45], the spectral-temporal-profile methodology of Foerster et al. [46], the inter-class separability index of Somers and Asner [47], and the Random Forest classifier of Breiman [48], with earlier dynamic-change analysis for Northeast China established by Huang et al. [49] and climate-suitability features supplied by the FAO land-evaluation framework [50] and the climatic-limitations approach of Holzkämper et al. [51]. Finally, adoption is not automatic: Long et al. [52] show that the business-model dimension of climate-smart agriculture diffusion in Europe (value proposition, channels, and key partnerships) is the principal bottleneck even under favourable subsidies, Pittelkow et al. [4] situate such interventions within the conservation-agriculture productivity-limits debate, and Waha et al. [8] argue for sowing-date and cropping-system choice as key climate-adaptation levers. Our solver builds directly on the deep-policy paradigm advanced by Mnih et al. [53] and the trust-region update of Schulman et al. [54], adapted to the discrete-binary action structure of parcel-level designation.

What is missing across these literatures is a single tool that operates at the parcel rather than the regional or irrigation-district level, that treats threshold-based site-selection compliance as a first-class

objective rather than a penalty term, and that delivers a designation plan directly consumable by a county GIS workflow. The cropping-structure studies diagnose the spatial mismatch but do not designate parcels; the metaheuristic and mathematical-programming toolboxes optimize flat, continuous, or aggregate encodings; and the econometric, remote-sensing, and diffusion strands supply the responsiveness, data, and adoption context but not the designation engine itself. In short, what is needed is a parcel-level, compliance-aware, GIS-ready designation tool that turns a band-shaped policy target into an executable map of which individual parcels to convert.

In this paper we therefore ask: given several thousand cultivated parcels, how should a county agricultural bureau allocate the strip-intercropping designation so as to (i) land within the policy area band, (ii) preserve wheat security, (iii) achieve high site-selection compliance, and (iv) maximize subsidy-augmented economic return? We approach the problem through reinforcement learning, specifically Proximal Policy Optimization [54], over a one-shot combinatorial designation Markov decision process (MDP). The state for each parcel encodes its size on two scales, its 2022 wheat–corn–soybean ratios, and its 2022 planting mode; the action is binary (convert or not); and the reward is composed at the episode level, encoding the four objectives above. Crucially, no parcel is removed from the decision pool for violating a site-selection rule: the three-class compliance score enters only through the reward, with an adjustable acceptance floor ρ^* (default 0.85), so compliant designation is a behaviour the policy must learn under an economic gradient that actively favours rule-violating parcels.

We instantiate the framework on Huojia County, Henan Province, one of the principal Huang–Huai–Hai cultivated-land regions, with a substantial wheat–corn double-cropping baseline, a designation as a national soybean science-and-technology self-reliance demonstration county, and a pledged 2023 expansion of strip intercropping from 3,000 to 10,000 mu under Henan's 1.5-million-mu provincial task. Our economic model integrates the 2023 subsidy stack: a 98 yuan/mu cultivated-land protection subsidy, a 25 yuan/mu grain-grower one-off subsidy, and a strip-intercropping specific subsidy specified by the 2023 provincial implementation plan as 150 yuan/mu of central finance plus a provincial top-up, which we model at a 200 yuan/mu total following county-level disclosures. With these layers in place, the per-mu profit ranking is strip intercropping (773 yuan/mu) > wheat–corn double cropping (348 yuan/mu) > single-season corn (251 yuan/mu) > corn–soybean intercropping (223 yuan/mu) > single-season wheat (220 yuan/mu) > wheat–soybean (194 yuan/mu) > single-season soybean (97 yuan/mu); note that the per-mu conversion gain to strip intercropping is therefore largest precisely for the modes whose parcels violate the isolated-location rule.

This paper demonstrates that a single large-scale reinforcement-learning formulation, trained over 8,833 simultaneous parcel decisions, can jointly satisfy the county-scale multi-objective constraint set (the area band, the wheat floor, an adjustable compliance floor, and economic improvement) and extend to a spatially coupled contiguity objective. The contributions are fourfold:

(1) A parcel-level reinforcement-learning (RL) formulation in which threshold-based site-selection compliance is learned from the reward, never imposed by pre-filtering: every rule-violating parcel stays in

the 8,833-parcel pool, yet the learned policy recovers all three rules from the reward alone (AUC ≈ 1.0, ρ = 0.871 across three seeds), whereas compliance-blind heuristics on the identical pool span ρ = 0.000 to a non-transferable 1.000 depending only on their sort key, so violation avoidance is a learned property of the policy, achieved against an economic gradient that favours violators, not an artefact of data pre-filtering.

(2) Compliance exposed as a tunable acceptance floor that a fixed-sort heuristic cannot offer: dialing ρ* moves the audited plan from 0.871 to 0.957 to exactly 1.000, the full-compliance plan coinciding with the size-sorted heuristic's plan at only 0.13 million yuan of forgone margin.

(3) A counterfactual (difference-reward) advantage estimator that repairs the credit-assignment collapse of one-shot PPO over 8,833 simultaneous decisions at O(|C|) cost, with a deterministic probability-ranked decoder, delivered as an end-to-end pipeline that emits GIS-ready shapefiles under a 2022–2023 subsidy-stack economic model.

(4) Evidence that the same reward template extends to a spatially coupled objective as a local change: a contiguity-aware variant reaches 76% versus 6% coherence at matched area against a spatially blind size-sorted heuristic, a coupling no fixed sort order can represent, at marginally lower (not higher) economics, so the supported axis is expressiveness rather than a benchmark win.

The paper is organized as follows. Section 2 details the data and implementation. Section 3 formalizes the problem setting and develops the MDP formulation and PPO solution method. Section 4 reports the empirical results on Huojia County. Section 5 presents the ablation studies. Section 6 discusses methodological choices, sensitivity considerations, and limitations, and Section 7 concludes.

2. Data and Implementation

2.1 Case study site

Huojia County, Henan Province, lies in the northern Huang–Huai–Hai Plain and is dominated by wheat–corn double cropping. The county is a national soybean science-and-technology self-reliance demonstration county and an active participant in the strip-intercropping program: the 2023 national program covered roughly 20 million mu, Henan's provincial task was 1.5 million mu (Yu Nong Wen (2023) No. 188), the Xinxiang municipal allocation was 116,000 mu, and Huojia planned to expand its program area from about 3,000 mu in 2022 to 10,000 mu in 2023. We encode this pledge as a new-designation band of 7,000 to 10,000 mu whose two ends carry exact policy meanings: if all of the 2022 program area re-enrolls, 7,000 mu of new designation completes the pledge; if none re-enrolls, 10,000 mu does. The band therefore hedges enrollment-renewal uncertainty rather than expressing an arbitrary tolerance. The wheat-area floor is set at 280,000 mu, anchored at approximately 95% of the county's observed wheat area. Two distinct ledgers must not be conflated here, and we keep them explicit throughout: the program ledger counts subsidy-enrolled area (about 3,000 mu in 2022; 10,000 mu pledged for 2023), whereas our pixel-based parcel layer measures co-planting practice and classifies 18,978 mu (211

parcels) as strip-like soybean–corn co-planting in 2022, about six times the enrolled area, for reasons Section 2.2 shows are expected properties of pixel-based mapping. The policy target is applied exclusively to newly designated area, and the strip-like class is used only to exclude already co-planted parcels from the decision pool. In particular, the plan never reduces existing co-planting: all newly designated parcels are drawn from land not currently co-planted, the 18,978 mu of strip-like parcels remain untouched outside the pool, and after planning the practice ledger rises to roughly 27,800 mu even as program enrollment meets the 10,000 mu pledge.

2.2 Parcel data

The input is a county-level parcel shapefile with attributes including total area (mu), 2022 wheat ratio, 2022 corn ratio, 2022 soybean ratio, and 2022 planting mode. Per-parcel crop fractions were obtained by overlaying three public pixel-level crop-distribution products on the parcel layer (the 30 m winter-wheat maps of Dong et al. [55], the 30 m soybean maps of Liu et al. [56], and the high-resolution single-season rice maps of Shen et al. [57]), computing, for each parcel, the fraction of its pixels carrying each crop signal; summer cropland pixels not attributed to soybean or rice were treated as corn. The 2022 planting mode was then assigned by two-level threshold rules on the per-parcel fractions, with a crop deemed present at ≥ 50% of parcel pixels and partially present at ≥ 30%: wheat ≥ 50% with corn ≥ 50% and soybean < 30% gives wheat–corn double cropping; wheat ≥ 50% with soybean ≥ 50% and corn < 30% gives wheat–soybean; wheat ≥ 50% with corn < 30% and soybean < 30% gives single-season wheat; wheat ≥ 50% with both corn ≥ 30% and soybean ≥ 30% (within-parcel co-occurrence of the two summer signals) is classified as strip-like co-planting (mode 6); analogous rules cover the non-wheat modes, and parcels satisfying no rule fall to the residual class. Two properties of this pixel-based provenance matter for what follows. First, the 2–4 m strips of strip intercropping are far below the 30 m pixel size, so the strip-like class is identified from within-parcel co-occurrence rather than from resolved strip geometry; mixed pixels, split planting within a cadastral parcel, and genuine non-enrolled co-planting all inflate it relative to the official program ledger (18,978 mu classified versus about 3,000 mu enrolled in 2022). We therefore treat the strip-like class as an upper-bound proxy for co-planting practice, never as a program quantity: the area objective is anchored to policy figures alone (Section 2.1), and the class is used only to exclude already co-planted parcels from designation. Second, the classification error propagates benignly into the remaining pipeline; Section 6.6 quantifies the only channel that touches a reported number, the baseline gross margin, and shows the conclusions are insensitive to it. The decision pool comprises every cultivated, non-strip parcel (2022 modes {1, 2, 3, 4, 5, 7}) with no compliance-based pre-filtering. Of the 17,833 county parcels, 8,833 (280,504 mu) form the pool; 5,381 of them (34,658 mu) violate at least one site-selection rule of Section 3.4, while 3,452 (245,847 mu) pass all three (component areas rounded independently). Violating parcels are retained and flagged, never removed: whether to avoid them is left entirely to the agent. Figure 1 shows the 2022 planting-mode distribution and Fig. 2 the pool composition.

2.3 Subsidy-augmented per-mu economic model

Per-mu profit is computed as the sum of subsidies calibrated to 2022–2023 Henan policy documents and the underlying agronomic net income reported by the Henan Provincial Price Cost Survey and Audit Bureau. Specifically: the cultivated-land protection subsidy is 98 yuan/mu, based on the 2022 public notices from neighbouring counties (Xiping 97.5 yuan/mu, Huaxian 97.48 yuan/mu); the grain-grower one-off subsidy is 25 yuan/mu, a conservative estimate within the 2022–2023 temporary subsidy range; the strip-intercropping specific subsidy is specified by the 2023 provincial implementation plan as 150 yuan/mu of central finance plus a provincial top-up of unstated size, which we model at a 200 yuan/mu total in line with county-level disclosures. The resulting per-mu profit by mode is reported in Table 1 (Section 3.3); under this schedule the 2022 baseline county gross margin reproduces from the Fig. 1 mode areas as the sum of area × profit = 110.62 million yuan, or 357.2 yuan/mu over the county's total cultivated area of 309,689 mu (the sum of all eight 2022 mode areas in Fig. 1). The 773 yuan/mu profit of strip intercropping is more than twice the wheat–corn double-cropping profit (Fig. 3); because the base 98 + 25 yuan/mu applies uniformly to modes 1–7, it cancels in every conversion delta, so results are insensitive to the proxied base components.

2.4 Software stack

The implementation is in Python 3.10 with NumPy 1.x and pandas; PPO is implemented directly in PyTorch (1.12) with the counterfactual-advantage estimator of Section 3.7, and GeoPandas/pyshp handle shapefile I/O. The same seed-management protocol as our complementary farm-scale framework is enforced. Mode-partitioned shapefile output uses ASCII filename mappings (e.g., 'wheat_corn.shp', 'strip_planting.shp') to avoid Windows filesystem encoding issues with Chinese characters, and the resulting directories are zipped for easy transfer to county GIS desktops.

3. Problem Setting and Method

3.1 Strip-intercropping policy context

The 2022–2023 Chinese central documents identify soybean–corn strip intercropping as a priority instrument, and implementation is cascaded through task allocations: roughly 20 million mu nationally in 2023, 1.5 million mu for Henan (Yu Nong Wen (2023) No. 188, with Appendix 1 allocating 116,000 mu to Xinxiang City), and county pledges beneath, Huojia's being an expansion from 3,000 to 10,000 mu. The provincial implementation plan promotes the 4:2 and 6:4 patterns (4–6 rows of soybean alternating with 2–4 rows of corn), specifies a 150 yuan/mu central-finance subsidy for enrolled area with a provincial top-up, requires village-level registration with public notice of at least seven days and spot checks covering no less than 10% of enrolled households, and bars a parcel from simultaneously receiving the cultivated-land rotation subsidy and the strip-intercropping subsidy. The intent is to narrow the per-mu profit gap with monoculture corn while preserving the rotational and biodiversity benefits of intercropping, building on the temporal-niche-differentiation evidence assembled by Yu et al. [15].

3.2 County-level operational requirements

County agricultural bureaus implementing the policy face four binding requirements. First, a newly designated strip-intercropping area within an upper- and lower-bounded band (7,000 to 10,000 mu in our Huojia County case study, the new designation that completes the county's 3,000-to-10,000 mu pledge under full versus zero renewal of the existing program area), sized to align with subsidy budgets and procurement capacity. Second, preservation of wheat acreage above a county-specific minimum (280,000 mu in our case study), ensuring continued participation in the national wheat procurement program. Third, site-selection compliance: each designated parcel should not have grown soybean continuously, should not be too small to be operationally manageable, and should not be spatially isolated from other strip-intercropping or wheat–corn parcels. Fourth, economic improvement, ensuring that aggregate gross margin at the county level is at least non-decreasing relative to the prior-year baseline.

3.3 Eight planting modes

The county's parcel-level dataset classifies each parcel into one of eight planting modes, summarized in Table 1.

Table 1
Eight planting modes with subsidy-augmented per-mu profit. Subsidy components: 98 yuan/mu cultivated-land protection (codes 1–7) + 25 yuan/mu grain-grower one-off (codes 1–7) + 200 yuan/mu strip-intercropping specific (code 6 only).

Code	Mode	Net income (yuan/mu)	Subsidy (yuan/mu)	Total profit (yuan/mu)
0	Other / fallow	0	0	0
1	Single-season wheat	97	123	220
2	Wheat–corn double cropping	225	123	348
3	Single-season corn	128	123	251
4	Wheat–soybean double	71	123	194
5	Single-season soybean	−26	123	97
6	Strip intercropping	450	323	773
7	Corn–soybean intercrop	100	123	223

3.4 Site-selection compliance taxonomy

We codify three operational compliance criteria into the reward function. Continuous-soybean violation: a parcel whose 2022 soybean ratio is at least 0.3 should not be designated for strip intercropping in 2023, to preserve rotational benefit and avoid disease accumulation. Small-parcel violation: a parcel with area below 20 mu is mechanically inefficient for strip intercropping and should not be designated. Isolated-location violation: a parcel whose 2022 mode is not in {2 (wheat–corn), 6 (strip)} is by construction

isolated from existing or adjacent strip-intercropping operations and should not be designated. The aggregate compliance rate is $\rho = m / D$, where D is the number of designated parcels and m the number of those parcels that are free of all three violations (2022 soybean ratio below 0.3, area at least 20 mu, and 2022 mode in {2, 6}). We use this all-or-nothing per-parcel test because a parcel is operationally acceptable for strip intercropping only if it violates none of the three rules, and we normalize by the designated count D rather than the whole county because compliance is a property of the chosen plan; ρ therefore reads directly as the share of designated sites that would pass an operational audit. The acceptance floor $\rho^* = 0.85$ is a framework default, not a government-mandated value: it requires the large majority of designated parcels to satisfy all three criteria while leaving the policy free to trade a small share of borderline parcels against area attainment where fully compliant parcels are scarce. A county can raise the floor, lower it, or switch to the zero-tolerance variant of Section 5.1 to match local regulation; Section 5.1 reports its effect on the learned behaviour. We emphasize that compliance is enforced nowhere else in the pipeline: the decision pool of Section 2.2 retains every violating parcel, so whatever compliance the designation achieves is a learned property of the policy, attributable to the reward term of Eq. (4) rather than to data preparation.

Having specified the policy setting, the operational requirements, the planting-mode economics, and the compliance taxonomy, we now develop the solution method. Figure 4 summarizes the end-to-end pipeline, from the parcel shapefile through decision-pool assembly, MDP encoding, and PPO training to the deterministic decoding and the mode-partitioned GIS output.

Three inputs, the parcel shapefile (area, 2022 wheat/corn/soybean ratios, 2022 mode), the subsidy-augmented economic model (150 yuan/mu central + provincial top-up, modelled at 200 yuan/mu, over base subsidies of 98 + 25 yuan/mu), and the policy constraints (7,000–10,000 mu new-designation band, 280,000 mu wheat floor, adjustable compliance floor ρ^*, default 0.85), feed a decision-pool assembly stage that retains all cultivated non-strip parcels (modes 1–5, 7), flagging rather than removing rule violators. The pool is encoded as an MDP with an 11-dimensional per-parcel state (two area scales, the 2022 crop ratios, a one-hot 2022-mode encoding; three neighbourhood features are appended in the contiguity variant), a binary designation action, and an episode-terminal additive reward. A multi-layer perceptron (MLP) Bernoulli policy is trained with PPO using per-parcel counterfactual (difference-reward) advantages and an annealed entropy bonus, then decoded deterministically by designation-probability ranking with a positive-marginal-reward acceptance test, producing the audited plan, planning CSVs, and mode-partitioned GIS shapefiles.

3.5 MDP formulation

We define a one-shot MDP $M = (S, A, P, r)$ over the decision pool (notation in Table 2); the episode consists of a single simultaneous designation step with a terminal reward, so a discount factor plays no role. Pool: parcel i enters the pool if its 2022 mode is a cultivated non-strip mode, $m_i \in \{1, 2, 3, 4, 5, 7\}$; no compliance criterion is applied at this stage. Let $C \subseteq \{1, \ldots, N\}$ denote the pool index set ($|C| = 8{,}833$ in our case study).

Table 2
Symbols used in the MDP formulation and reward design.

Symbol	Description
$M = (S, A, P, r)$	one-shot Markov decision process: state space, action space, transition, terminal reward
C, N	decision-pool set (cultivated non-strip modes {1, 2, 3, 4, 5, 7}; no compliance pre-filter) and total county parcel count
s_i	state vector of pool parcel i
a_i	area of parcel i (mu)
a_{max}	maximum parcel area in the county (mu)
w_i, c_i, b_i	2022 wheat, corn, and soybean area ratios of parcel i
ℓ_i	logarithmic area scale $\log(1 + a_i) / \log(1 + a_{max})$, preserving resolution among small parcels
m_i	one-hot encoding of the 2022 planting mode of parcel i over the pool modes {1, 2, 3, 4, 5, 7}
$A(j, i)$	counterfactual (difference-reward) advantage of decision i in sampled plan j (Section 3.7)
u	action vector; component $u_i = 1$ designates parcel i for strip intercropping, 0 keeps it at its 2022 mode
A_{strip}	newly designated strip-intercropping area of the plan (mu)
A_{min}, A_{max}	lower and upper bounds of the new-designation band (7,000 and 10,000 mu)
A_{mid}	midpoint of the new-designation band (8,500 mu)
A_{wheat}, A_{wmin}	wheat-security area of the plan and its county floor (280,000 mu)
ρ, ρ^*	achieved site-selection compliance rate and its acceptance floor (framework default 0.85, adjustable)
Π, Π_{prev}	subsidy-augmented gross margin of the plan and of the 2022 baseline (yuan)
R_{area}, $R_{compliance}$, R_{wheat}, $R_{economic}$	area, compliance, wheat-security, and economic reward terms
R_{cont}, w_{cont}, w_{clust}	spatial-contiguity reward term and its coherence and cluster weights
φ, κ	spatial-coherence fraction and number of connected designated clusters

For each parcel $i \in C$ the per-parcel state is the eleven-dimensional vector:

$s_i = [\, a_i / a_{max},\ \ell_i,\ w_i,\ c_i,\ b_i,\ m_i \,]$ (1)

where a_i is the parcel area, a_{max} the maximum parcel area, $\ell_i = \log(1 + a_i) / \log(1 + a_{max})$ a logarithmic area scale that preserves resolution among small parcels (the small-parcel rule binds at 20 mu while a_{max} = 432 mu), (w_i, c_i, b_i) the 2022 wheat, corn, and soybean ratios, and m_i the one-hot encoding of the 2022 planting mode over the six pool modes. The state deliberately contains no violation flag: the three compliance rules are threshold functions of (b_i, a_i, m_i) that the policy must discover from the reward signal.

The action is a vector $u \in \{0, 1\}^{|C|}$, with $u_i = 1$ indicating designation of parcel i for strip intercropping; all |C| decisions are taken in a single episode, a one-shot combinatorial structure whose credit-assignment consequences are addressed in Section 3.7.

Transition. Deterministic: applying u maps pool parcel i to mode 6 if $u_i = 1$ and leaves it at its 2022 mode if $u_i = 0$, with non-pool parcels unchanged. Mode 6 denotes winter wheat followed by summer soybean–corn strips: a conversion from wheat–corn (mode 2) replaces only the summer corn season, leaving parcel wheat area invariant, while conversions from the summer-only modes {3, 5, 7} add a winter wheat season.

The reward is evaluated at the episode terminus from the resulting plan and decomposes additively into an area, a compliance, a wheat-security, and an economic term:

$r = R_{area} + R_{compliance} + R_{wheat} + R_{economic}$ (2)

The area term is band-shaped: it is positive inside the policy band, peaks at the band midpoint, and decays toward the band edges,

$R_{area} = 50 + 20\,(1 - |A_{strip} - A_{mid}| / A_{mid}),\ A_{min} \le A_{strip} \le A_{max}$ (3)

while outside the band it applies a linear shortfall penalty $-30\,(A_{min} - A_{strip}) / A_{min}$ below the lower bound and a steeper overshoot penalty $-100\,(A_{strip} - A_{max}) / A_{max} - 50$ above the upper bound. The compliance term rewards a plan whose site-selection compliance rate meets the acceptance floor ρ* and penalizes any shortfall,

$R_{compliance} = +20$ if $\rho \ge \rho^*$; $-50\,(\rho^* - \rho) / \rho^*$ otherwise (4)

the wheat-security term penalizes any wheat area below the county floor,

$R_{wheat} = -10\,(A_{wmin} - A_{wheat}) / A_{wmin},\ A_{wheat} < A_{wmin}$ (5)

and the economic term, active only inside the area band, scales the subsidy-augmented gross-margin improvement:

$R_{economic} = 5\,(\Pi - \Pi_{prev}) / 10^6$, $A_{min} \le A_{strip} \le A_{max}$ (6)

All symbols are defined in the Table 2 above.

The band-shaped area reward is a key design element: it encodes the policy reality that under-attainment and over-attainment are both costly, while leaving the within-band optimum to the interplay of objectives. At the real county scale the two within-band slopes nearly cancel: the economic slope ($5 \times 425 / 10^6 \approx 0.0021$ per mu) and the area term's decay toward the band edges ($20 / 8{,}500 \approx 0.0024$ per mu) leave the composite reward almost flat across the upper half of the band, and the trained policies land between 8,600 and 10,000 mu accordingly (Section 4). The sharp out-of-band penalties act as guardrails (Section 5.3 shows a size-sorted heuristic breaching the ceiling that the learned policies respect), and a regulator preferring a particular landing point can re-weight the area term; this trade-off is part of the tunable policy interface (Section 6.2).

To capture the operational value of spatially coherent designations, the framework supports an optional spatial-contiguity objective. We build a parcel-adjacency graph by linking pool parcels whose centroids lie within 500 m (70,269 edges over the 8,833-parcel pool), extend the per-parcel state with three static neighbourhood features (neighbour count, total neighbour area, and the compliant share of neighbours), and add a fifth reward term $R_{cont} = w_{cont} \cdot \varphi - w_{clust} \cdot \kappa$, where φ is the spatial-coherence fraction (the share of designated parcels with at least one designated neighbour) and κ is the number of connected designated clusters; we use $w_{cont} = 60$ and $w_{clust} = 1.0$, with weight sensitivity, and a structural upper limit on w_{clust}, reported in Section 5.3. The full reward becomes $r = R_{area} + R_{compliance} + R_{wheat} + R_{economic} + R_{cont}$; the counterfactual advantages of Section 3.7 extend to the contiguity terms through local graph deltas, and the decoder maintains a union–find structure so its marginal-reward test is exact. Its effect is reported in Section 5.3.

3.6 Heuristic baselines

For comparison we implement deterministic compliance-blind heuristics of the family a county officer might apply: sort the pool by a key, then select from the top until a target area is reached. The size-sorted greedy uses descending parcel area ('designate the largest parcels first to hit the target with the fewest entries'). The profit-sorted greedy uses descending per-mu conversion gain, the economically rational key, since converting low-revenue modes gains the most per mu (Table 1: 676 yuan/mu for single-season soybean versus 425 for wheat–corn). Neither key encodes compliance, and their realized compliance on the same pool differs dramatically (Section 4): every parcel the profit key selects violates a rule ($\rho = 0.000$), while the size key happens to attain $\rho = 1.000$ on Huojia only because this county's violating parcels are small (mean 6.4 mu), a data accident, not a property of the method. We additionally report a random-ranking control for the decoding step of Section 3.7, which isolates how much of the final-plan quality is contributed by the learned probabilities rather than by the decoder.

3.7 PPO with counterfactual advantages

We use a multi-layer perceptron (MLP) policy (two hidden layers of 64 tanh units) mapping the 11-dimensional parcel state to one Bernoulli logit per parcel. PPO hyperparameters follow Schulman et al. [54]: clip parameter 0.2, learning rate 3×10^{-4}, 10 update epochs per batch, and an entropy coefficient initialized at 0.01 and annealed linearly to zero over the first 70% of training; each update samples K = 16 plans from the current policy, and training runs for 1,200 updates.

A naive application of PPO to the one-shot designation structure fails. With a single episode-level reward shared by |C| = 8,833 simultaneous Bernoulli decisions, every decision receives the same advantage, per-parcel credit assignment vanishes by averaging, and in our experiments the policy degenerates into hoarding small rule-violating parcels for fine-grained area control, with compliance collapsing toward zero. We therefore assign each decision its counterfactual (difference-reward) advantage $A(j, i) = R(u^{(j)}) - R(\text{flip}_i(u^{(j)}))$, the change in episode reward if decision i alone were reversed [58, 59]. Because every reward term depends on the plan only through five aggregates (designated area, gross-margin change, wheat area, designated count, violation count), all |C| counterfactuals of a sampled plan are computed vectorially at O(|C|) cost. Clipped-surrogate updates then proceed exactly as in standard PPO, with A(j, i) in place of a shared advantage.

Deterministic evaluation requires care for the same structural reason: interchangeable large compliant parcels leave the optimal stochastic policy mixed (no single parcel's probability need exceed 0.5), so naive probability thresholding decodes an empty plan. We instead decode by the learned ranking: parcels are visited in descending designation probability and a parcel is kept only if its marginal reward contribution is positive. The decoder is deterministic and reproducible; Section 4 quantifies, via the random-ranking control, that the decoder's marginal-reward test holds the compliance rate near the floor under any ranking (a random ranking already reaches ρ = 0.854), while the learned ranking contributes the economic and parcel-quality gains and the near-complete avoidance of violators (AUC ≈ 1.0).

3.8 Determinism guarantees

The environment is fully deterministic given the action vector; reproducibility is enforced through a single global seed propagated across NumPy and PyTorch. Given a published seed and a fixed software environment, the full pipeline is deterministic and reproducible. Across three independent seeds, the zero-tolerance variant decodes to an identical plan, and the default-reward plans coincide on all aggregate metrics to within about 2.5% (8,828–9,010 mu; ρ = 0.850–0.871) while differing modestly in parcel composition; we report seed ranges alongside the headline plan. This dual guarantee (exact replicability of the artefact, demonstrated stability of the method) is essential for a tool intended to inform actual subsidy disbursement.

3.9 Outputs

The pipeline produces three artefacts. First, a parcel-level CSV with 2022 mode, 2023 plan mode, conversion indicator, and per-parcel revenue change. Second, a planning-summary CSV listing aggregate metrics (total revenue, designated area, compliance rate, wheat area). Third, a set of GIS shapefiles

partitioned by 2022 and 2023 modes and a separately exported new-converted-strip shapefile containing only the newly designated parcels, directly consumable in county GIS platforms such as ArcGIS or QGIS.

4. Results

We trained the policy on the full Huojia County decision pool (8,833 parcels, 280,504 mu, of which 5,381 parcels violate at least one site-selection rule) for 1,200 updates under the reward of Eqs. (2)–(6) with the new-designation band of 7,000–10,000 mu, repeating the experiment with three random seeds. Decoding the converged policies yields closely agreeing plans (8,828 / 9,010 / 8,864 mu; 31 / 40 / 31 parcels; ρ = 0.871 / 0.850 / 0.871); we report seed 0 as the published plan. It designates 31 parcels totalling 8,828 mu of newly converted strip intercropping: 27 large wheat–corn parcels (mean 318.6 mu, minimum 263.6 mu) plus four rule-violating parcels discussed below, fulfilling the county's pledged expansion within the 7,000–10,000 mu band (Fig. 5a, Fig. 6). Wheat-security area rises slightly to 295,628 mu; three of the four violating conversions are from summer-only modes and add a winter wheat season; the wheat–soybean parcel already carries one, above the 280,000 mu floor (Fig. 5d). The wheat-security term is non-binding on Huojia: the baseline already clears the floor by some 15,000 mu and no conversion in the pool can reduce wheat area, so here it functions as a safeguard rather than an active constraint. The subsidy-augmented county gross margin rises from 110.615 to 114.403 million yuan, an increase of 3.79 million yuan (+ 3.4%), with per-mu return improving from 357.2 to 369.4 yuan/mu (Fig. 5b, c). Table 3 details the 2022-to-2023 changes and Table 4 the requirement-by-requirement audit.

A parcel-level audit of the designated set gives ρ = 0.871: 27 of 31 parcels pass all three rules, and the plan sits just above the default acceptance floor ρ* = 0.85. The four exceptions (a 15.6 mu and a 27.4 mu single-season-soybean parcel, a 74.1 mu wheat–soybean parcel, and a 109.8 mu corn–soybean parcel; 227 mu combined) are precisely the highest per-mu-gain conversions in the pool, and all four carry learned designation probabilities below 0.001, ranked at the very bottom by the policy itself; they enter through the decoder's marginal-reward test, which prices their inclusion at zero compliance cost while ρ remains above the floor. This is the designed behaviour of an acceptance-floor compliance term: the framework spends the tolerance the regulator grants. Section 5.1 shows that the spent tolerance follows the dial: raising ρ* to 0.95 lifts the audited plan to 0.957, and a zero-tolerance fine eliminates all four parcels and attains ρ = 1.000 exactly.

Figure 7 shows the training dynamics behind these plans: sampled-plan compliance starts near the pool composition (ρ ≈ 0.3–0.5), rises above the 0.85 floor within roughly 200 updates, and stabilizes just above it, while the newly designated area settles into the band. The designation probabilities of the converged policies separate compliant from violating parcels essentially perfectly: the mean learned probability is ≈ 0.007 for violating versus ≈ 0.08 for compliant parcels, with AUC between 0.9998 and 1.0000 across seeds; the policy has learned to avoid rule-violating parcels from the reward signal alone (the three rule thresholds are functions of features already in the state, so high separability is expected;

the substantive result is that compliance is achieved against an adverse economic gradient), without ever observing a violation flag.

Table 5 reports the learned plan alongside compliance-blind heuristics that serve as controls, illustrating the problem structure and confirming that the achieved compliance is a property of the policy rather than of a sort key, all evaluated on the same unfiltered pool at the same task. The profit-sorted greedy, the economically rational rule, designates 476 parcels at ρ = 0.000: every single parcel it selects violates a rule, because the per-mu conversion gain is maximal for the violating modes; it nonetheless posts the largest raw margin gain, which is exactly the temptation the compliance term must defeat. The size-sorted greedy attains ρ = 1.000 on this county only by the data accident that violating parcels are small. The random-ranking decode control reaches ρ = 0.854, the floor the decoder is given, while the learned policy delivers ρ = 0.871 with a 3.6% larger margin gain than the size-sorted plan; this gap is spent compliance tolerance, not a structural edge, because under the zero-tolerance setting the policy collapses to the size-sorted heuristic's identical 8,601 mu / 27-parcel plan, the extra margin being only the four high-gain violating parcels the granted floor admits. The advantage of the RL formulation here is not raw economics (a size-sorted heuristic handed the correct target area ties it) but that band position, parcel choice, and compliance margin all emerge jointly from the reward specification, with no hand-supplied target, no sort key, and no dependence on the pool happening to be benign; Section 5.3 shows the further regime, spatial contiguity, where fixed sort orders fail outright.

Table 3
Huojia County strip-intercropping plan: 2022 baseline versus 2023 PPO designation.

Metric	**2022 baseline**	**2023 PPO plan**	**Change**
Newly designated strip area (mu)	—	8,828	fulfils 3,000 → 10,000 pledge
Newly designated parcels	—	31	mean 284.8 mu/parcel
Total gross margin (10^4 yuan)	11,061.5	11,440.3	+ 378.8 (+ 3.4%)
Per-mu return (yuan/mu)	357.2	369.4	+ 12.2
Wheat-security area (mu)	295,475	295,628	+ 153 (added winter-wheat season)
Site-selection compliance	—	0.871 (27 of 31 compliant)	learned; AUC ≈ 1.0

Table 4
Compliance of the PPO plan with the four operational requirements.

Operational requirement	Target	Achieved	Status
Strip-intercropping area	7,000 to 10,000 mu new designation	8,828 mu	Pass
Wheat-security area	≥ 280,000 mu	295,628 mu	Pass
Site-selection compliance	≥ 0.85 (framework default, adjustable)	0.871 (27 of 31 compliant)	Pass
Economic improvement	≥ 2022 baseline	+ 3.4% (+ 3.79M yuan)	Pass

Table 5
Compliance ladder on the unfiltered decision pool. The greedy heuristics are run at the 8,500 mu mid-band target they require as an input; the random-ranking decode control and the learned policy determine their own in-band position. The control applies the Section 3.7 decoder to a random parcel ranking, isolating the decoder's contribution from the learned probabilities.

Plan	Compliance ρ (violations/parcels)	New designated area (mu)	Gross-margin gain (10^4 yuan)
Profit-sorted greedy (compliance-blind)	0.000 (476/476)	8,503	471.3
Size-sorted greedy (compliance-blind)	1.000 (0/27)	8,601	365.5
Random-ranking decode (control)	0.854 (20/137)	8,515	362.0
PPO (learned policy; seed 0 of 3)	0.871 (4/31)	8,828	378.8

5. Ablation Studies

We organize the ablation experiments into three families: sensitivity to the operational thresholds and the reward design (Section 5.1), the subsidy counterfactual (Section 5.2), and the algorithm-design choices (Section 5.3). The threshold and reward-design analyses and the contiguity comparison are executed on the real parcel data; the subsidy retraining and the network-architecture comparison are planned analyses that have not yet been run (Section 5.2 and the first part of Section 5.3).

5.1 Sensitivity to operational thresholds and reward design

The acceptance floor ρ* is a tunable interface, and its effect on the learned behaviour is empirical. Under the default threshold-shaped compliance term (Eq. 4) the audited plan lands just above the floor (ρ = 0.871 for ρ* = 0.85): once above ρ*, the marginal compliance penalty is zero, so the policy spends part of the granted tolerance on the highest-gain violating parcels; an acceptance floor behaves, as designed, as

a floor rather than a target. Dialing the floor moves the behaviour with it: retraining with ρ* = 0.95 lifts the audited plan to ρ = 0.957 (two violating parcels instead of four), and a zero-tolerance variant that adds a constant per-violation fine attains ρ = 1.000 exactly: zero violating parcels, an identical 8,601 mu / 27-parcel plan from all three seeds, at a cost of only 13.3 $\times 10^4$ yuan (0.13 million yuan) of forgone margin gain (365.5 versus 378.8 $\times 10^4$ yuan). The compliance level is therefore a controllable function of the reward design, not an artefact of the data. The two designs also separate under scarcity. Raising the small-parcel rule to 300 mu shrinks compliant supply to 5,542 mu, below the 7,000 mu band floor, making full compliance and band attainment mathematically incompatible; the zero-tolerance configuration then designates exactly the 16 available compliant parcels and stops short of the band, surfacing the infeasibility conservatively instead of fabricating a violating plan. (We note candidly that this conservative fixed point is not the reward optimum, which would enter the band with minimal violations; escaping it requires stronger exploration than our default protocol.) The continuous-soybean criterion binds on only 12 pool parcels in Huojia, so its 0.3 sub-threshold is retained for transferability rather than swept.

The operationally decisive data-side lever is the minimum-parcel-area rule. Figure 8 sweeps it from 10 to 400 mu against the compliant supply: the 10,000 mu ceiling target stays attainable up to a rule of about 260 mu and the 7,000 mu band floor up to about 275 mu, beyond which the pledge becomes infeasible for fully compliant designation, the stress regime examined above. Together these sweeps give provincial regulators a quantitative basis for setting the minimum-parcel-area requirement, the strictness of the compliance term, and the sharpness of the area target.

5.2 Subsidy counterfactual

The 200-yuan/mu strip-intercropping specific subsidy is the dominant economic driver. As a planned analysis that has not yet been run, the agent would be retrained at a sequence of specific-subsidy levels {0, 50, 100, 150, 200, 250, 300} yuan/mu, holding all other parameters fixed. At zero specific subsidy the strip-intercropping per-mu profit (573 yuan/mu) still exceeds the wheat–corn double-cropping profit (348 yuan/mu); we anticipate that the agent would still designate parcels but with a smaller margin of preference, so the designated area would contract toward the lower bound of the band. The compliance level attained at each subsidy step would itself be a primary outcome: a smaller strip premium changes the relative temptation of rule-violating parcels, and the sweep would quantify how robustly the threshold-shaped term holds ρ above the floor. At 300 yuan/mu the strip-intercropping advantage would be so dominant that the area-band ceiling would become the binding constraint. Figure 9 reports only the underlying per-mu economic model that motivates this analysis (the relative profit of strip intercropping versus wheat-corn double cropping across the subsidy range) and is not itself a trained-policy result; the policy-cost-effectiveness curve of subsidy versus designated area is left to this future work.

5.3 Algorithm design: architecture and baseline

We probe two algorithmic design choices. First, network capacity: a planned comparison (not yet run) of policy MLPs of 64×64, 128×128, and 256×256 under otherwise identical PPO hyperparameters, where we anticipate modest differences because the per-parcel state is low-dimensional, with a preference for the smaller networks on training-cost grounds. Second, spatial contiguity, the regime in which fixed sort orders fail structurally. We train the contiguity variant of Section 3.5 (neighbourhood-extended state; w_{cont} = 60, w_{clust} = 1.0) on the same unfiltered pool, so compliance and contiguity must be learned jointly, and compare against the size-sorted greedy heuristic at matched new area. As Fig. 10 and Table 6 show, at approximately 10,000 mu of new designation the greedy heuristic scatters 33 parcels into 32 clusters at 6.1% spatial coherence, and its prefix selection overshoots the band ceiling (10,174 mu), whereas the contiguity-aware policy concentrates 42 parcels into 23 clusters at 76.2% coherence, exactly inside the band (10,000 mu), at ρ = 0.881 and near-identical economics (425.0 versus 432.4 ×10^4 yuan). Coherence rises with the contiguity weights (56.8% at w_{cont} = 40, w_{clust} = 0.5; 76.2% at 60 and 1.0), and the weights face a structural limit: the per-cluster penalty must stay below the band-floor area marginal of a seed parcel, or no first parcel can open a cluster; at w_{clust} well above 1 the policy collapses toward empty plans, and one strong-weight seed indeed converged to a conservative sub-band fixed point (4,605 mu at perfect coherence), echoing the exploration caveat of Section 5.1. We also note the geometric ceiling honestly: at the real county scale a plan comprises only a few dozen parcels drawn from a scattered compliant supply, so coherence cannot approach the near-total values attainable when many small parcels are designated; the salient contrast is the order-of-magnitude gap to the sort-based heuristic at matched area, a spatial coupling no fixed sort order can express. We do not claim this beats a purpose-built spatial heuristic; the policy's economics here are in fact marginally lower, not higher, and a cluster-aware seed-and-grow comparator remains future work, so the supported claim is expressiveness: one reward formulation encodes a neighbour-coupled objective as a local change that a fixed sort order cannot.

Table 6
Spatial-contiguity comparison of the contiguity-aware PPO policy and the size-sorted greedy heuristic at matched new area on the unfiltered pool.

Metric	Size-sorted greedy	Contiguity-aware PPO
Designated area, new (mu)	10,174 (exceeds band ceiling)	10,000 (inside band)
Gross-margin improvement (10^4 yuan)	432.4	425.0
Site-selection compliance	1.000 (0/33)	0.881 (5/42)
Designated parcels	33	42
Spatial clusters	32	23
Spatial coherence (≤ 500 m)	6.1%	76.2%

6. Discussion

6.1 Rationale for an RL formulation

The combinatorial site-selection structure of the problem is awkward to encode in a single linear or quadratic program, while greedy heuristics, though tractable, bind their behaviour to a hand-chosen sort key and a hand-supplied target. We are explicit about where reinforcement learning is (and is not) doing the work. On economics alone, a size-sorted heuristic handed the correct target ties the learned policy. The learned policy's distinctive contributions on the base objectives are that compliance, band position, and parcel choice emerge jointly from the reward specification: compliance is learned against an adverse economic gradient and is robust to pool composition (Section 4), where heuristic compliance ranges from 0.000 to 1.0 depending on the sort key, and the learned compliance level tracks the regulator's dial (Section 5.1). The formulation's distinctive contribution appears once a spatially coupled objective such as contiguity is introduced (Section 5.3), where the marginal value of designating a parcel depends on its neighbours' designations, a coupling no fixed sort order can express; we anticipate that a purpose-built cluster-aware heuristic (for example, seed-and-grow) could close part of the coherence gap, and we commit to that baseline in the program-year revision. Our claim is therefore deliberately narrower than 'PPO beats heuristics': each newly coupled policy objective forces a bespoke redesign of a heuristic, whereas the MDP formulation absorbed compliance learning, the contiguity objective, and the credit-assignment repair of Section 3.7 as local changes to the reward and the estimator, with the same absorption path open for the multi-year rotation and uncertainty extensions of Section 6.6. A systematic head-to-head comparison against other planning models, such as mathematical-programming and metaheuristic optimizers, is left to future work.

6.2 The role of the band-shaped area reward

Standard linear-programming objectives tend to push the designated area toward one end of the feasible band, because the linear coefficient on area is constant within the band. The band-shaped reward we adopt (strictly positive inside the band, with an area term maximal at the midpoint, and sharply negative outside) replaces that arbitrariness with an explicit design: the out-of-band cliffs are guardrails (which the size-sorted heuristic of Section 5.3 breaches while the learned policies do not), and the within-band landing point is set by the interplay between the area and economic terms, which at the real county scale nearly cancel, leaving the policy free to land anywhere in the upper band. A regulator who prefers a particular landing point simply re-weights the area term. Section 5.1 quantifies the data-side sensitivity of band attainability.

6.3 Compliance taxonomy as policy interface

Encoding the three violation types as separately tracked counters, rather than collapsing them into a single composite penalty, is a deliberate design choice. It allows county officers to reason about which kind of violation is binding in a given plan and to revise policy parameters accordingly. For instance, if continuous-soybean violations dominate, the threshold ratio (currently 0.3) can be tightened; if isolated-location violations dominate, the permissible-mode set can be expanded to include wheat–soybean parcels. The acceptance floor ρ* itself belongs to this interface, and Section 5.1 demonstrates it

experimentally: the audited compliance of the learned plan tracks the dial from 0.871 at the default floor, to 0.957 at ρ* = 0.95, to 1.000 under the zero-tolerance variant. The transparency of this decomposition aligns with the broader call for interpretable RL in high-stakes operational contexts.

6.4 Subsidy sensitivity and counterfactual analysis

Section 5.2's subsidy counterfactual would, once run, convert the same trained policy into a policy-cost-effectiveness curve. Such analysis can inform future revisions of the subsidy schedule and provide quantitative ammunition in budget negotiations, complementing the equilibrium-style analyses of Zheng et al. [38] on producer responses to price reform and the broader supply-elasticity literature [39, 41].

6.5 Generalization to other counties and provinces

The architecture is intentionally county-agnostic: changing the input shapefile and updating the per-mu profit table is sufficient to retarget the framework to a different county. For provinces other than Henan the subsidy structure may need adjustment (different specific-subsidy levels, different cultivated-land protection rates), but the reward-function template is unchanged. We expect the framework to be most directly transferable to other Huang–Huai–Hai counties with comparable wheat–corn double-cropping baselines (Hebei, Shandong, Anhui), and to require modest reward re-weighting for transfer to Northeast provinces where single-season cropping prevails [3, 38].

6.6 Limitations

First, the framework optimizes a single-year designation and does not encode multi-year rotation dynamics; this is acceptable in the current pilot stage but should be revisited as multi-year program data accumulate. Second, the spatial-isolation compliance criterion is captured through a categorical mode indicator; the spatial-contiguity objective of Section 5.3 takes a first step beyond this by building a centroid-distance neighbourhood graph, but a fuller treatment would use true polygon adjacency and could fold contiguity directly into the compliance taxonomy. Third, the economic parameters are deterministic; uncertainty in realized yields, market prices, and subsidy disbursement rates could be added through scenario-based or distributional reward formulations. Fourth, the one-shot combinatorial-bandit formulation, while honest to the decision structure, required the counterfactual-advantage estimator of Section 3.7 and a ranking-based decoder; a sequential per-parcel formulation with explicit progress features is a natural alternative under which standard advantage estimators apply. Fifth, under severe compliant-parcel scarcity the zero-tolerance configuration converges to a conservative fixed point short of the reward optimum (Section 5.1); stronger exploration or curriculum schedules are needed before deploying that variant in scarce regimes. Finally, the parcel attributes derive from public pixel-level crop maps (Section 2.2) whose strip-like baseline class is an upper-bound proxy: in the extreme case that all 18,978 mu of strip-like parcels were in fact conventional wheat–corn, the 2022 baseline gross margin would fall by 8.07 million yuan (to 102.55 million), raising the plan's relative improvement from + 3.4% to + 3.7% while leaving the designation itself unchanged; the plan never designates those parcels and the area objective never references them, so the classification uncertainty does not propagate into the plan.

7. Conclusion

We have described a parcel-level Proximal Policy Optimization framework for county-scale soybean–corn strip intercropping designation, instantiated on Huojia County, Henan Province, against the county's real 2023 program pledge (expansion from 3,000 to 10,000 mu under Henan's 1.5-million-mu provincial task). Distinctively, the framework imposes site-selection compliance nowhere in data preparation: every rule-violating parcel remains in the decision pool, and compliance, defined by the auditable rate $\rho = m / D$ over the designated set, with an adjustable acceptance floor ρ^* (default 0.85), is learned from the reward signal against an economic gradient that favours violations. On the full county pool the learned policy designates 31 parcels (8,828 mu of new strips) inside the encoded 7,000–10,000 mu band, slightly raises wheat area, improves the subsidy-augmented gross margin by 3.79 million yuan (+ 3.4%) over the 2022 baseline, and audits at $\rho = 0.871$, against 0.000 for an economically rational compliance-blind heuristic; the compliance level tracks the regulator's dial, reaching 0.957 at $\rho^* = 0.95$ and exactly 1.000 under a zero-tolerance fine, and a neighbourhood-conditioned contiguity variant achieves more than an order of magnitude higher spatial coherence than a size-sorted heuristic at matched area (76.2% versus 6.1%). The framework consumes a parcel shapefile and produces planning tables and mode-partitioned GIS shapefiles directly usable in county-level platforms. Three directions stand out for future work: (i) multi-year program planning that captures rotational accumulation; (ii) explicit polygon-adjacency modelling, extending the contiguity results of Section 5.3; and (iii) provincial-scale aggregation, in which the same framework is run across all participating counties to support province-wide subsidy-budget allocation. The subsidy-retraining and architecture analyses outlined in Section 5 are planned as future work.

Declarations

Acknowledgements

This research was supported in part by the National Key Research and Development Project of China under Grant 2023YFD2000105-02.

References

1. Li T (2023) Planting structure adjustment and layout optimization of feed grain and food grain in China based on productive potentials. Land 12(1):45
2. Zhan S (2022) The political economy of food import and self-reliance in China: 1949–2019. Global Food History 8:194–212
3. Liu Z, Yang P, Wu W et al (2018) Spatiotemporal changes of cropping structure in China during 1980–2011. J Geog Sci 28(11):1659–1671
4. Pittelkow CM, Liang X, Linquist BA et al (2015) Productivity limits and potentials of the principles of conservation agriculture. Nature 517:365–368

5. Wu W, Yu Q, You L et al (2018) Global cropping intensity gaps: Increasing food production without cropland expansion. Land Use Policy 76:515–525
6. Tang H, Wu W, Yang P et al (2010) Recent progresses in monitoring crop spatial patterns by using remote sensing technologies. Scientia Agricultura Sinica 43(14):2879–2888 (in Chinese)
7. Liu Z, Yang X, Chen F et al (2013) The effects of past climate change on the northern limits of maize planting in Northeast China. Clim Change 117(4):891–902
8. Waha K, Müller C, Bondeau A et al (2013) Adaptation to climate change through the choice of cropping system and sowing date in sub-Saharan Africa. Glob Environ Change 23(1):130–143
9. Aguilar J, Gramig GG, Hendrickson JR et al (2015) Crop species diversity changes in the United States: 1978–2012. PLoS ONE 10:e0136580
10. Hijmans RJ, Choe H, Perlman J (2016) Spatiotemporal patterns of field crop diversity in the United States, 1870–2012, vol 1. Agricultural & Environmental Letters, p 160022
11. Leff B, Ramankutty N, Foley JA (2004) Geographic distribution of major crops across the world. Glob Biogeochem Cycles 18:GB1009
12. Zhang G, Xiao X, Biradar CM et al (2017) Spatiotemporal patterns of paddy rice croplands in China and India from 2000 to 2015. Sci Total Environ 579:82–92
13. Fischer G, Nachtergaele FO, van Velthuizen HT et al (2021) Global Agro-Ecological Zones v4: Model Documentation. FAO
14. Song G, Zhang H (2021) Cultivated land use layout adjustment based on crop planting suitability: A case study of typical counties in Northeast China. Land 10(2):107
15. Yu Y, Stomph TJ, Makowski D et al (2015) Temporal niche differentiation increases the land equivalent ratio of annual intercrops: A meta-analysis. Field Crops Res 184:133–144
16. Jain R, Malangmeih L, Raju SS et al (2018) Optimization techniques for crop planning: A review. Indian J Agric Sci 88(12):1826–1835
17. Kaur B, Sidhu RS, Vatta K (2010) Optimal crop plans for sustainable water use in Punjab. Agricultural Econ Res Rev 23(2):273–284
18. Mainuddin M, Das Gupta A, Onta PR (1997) Optimal crop planning model for an existing groundwater irrigation project in Thailand. Agric Water Manage 33(1):43–62
19. Mohaddes SA, Mohayidin MG (2008) Application of the fuzzy approach for agricultural production planning in a watershed. American-Eurasian J Agric Environ Sci 3(4):636–648
20. Norton RD, Schiefer GW (1980) Agricultural sector programming models: A review of alternative approaches. Eur Rev Agric Econ 7:229–265
21. Sarker RA, Talukdar S, Haque AFMA (1997) Determination of optimum crop mix for crop cultivation in Bangladesh. Appl Math Model 21(10):621–632
22. Sethi LN, Panda SN, Nayak MK (2006) Optimal crop planning and water resources allocation in a coastal groundwater basin, Orissa, India. Agric Water Manage 83(3):209–220

23. Sharma DK, Jana RK, Gaur A (2007) Fuzzy goal programming for agricultural land allocation problems. Yugoslav J Oper Res 17(1):31–42
24. Kennedy J, Eberhart R (1995) Particle swarm optimization. Proceedings of ICNN'95, International Conference on Neural Networks, 4, 1942–1948
25. Coello-Coello CA, Pulido GT, Lechuga MS (2004) Handling multiple objectives with particle swarm optimization. IEEE Trans Evol Comput 8(3):256–279
26. Begam S, Jain R, Arora A et al (2023) Multi-objective particle swarm optimization for regional crop planning. Indian J Agric Sci 93(2):237–240
27. Deb K, Pratap A, Agarwal S et al (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197
28. Sarker R, Ray T (2009) An improved evolutionary algorithm for solving multi-objective crop planning models. Comput Electron Agric 68:191–199
29. Chen Y, Zhou Y, Fang S et al (2022) Crop pattern optimization for the coordination between economy and environment considering hydrological uncertainty. Sci Total Environ 809:151152
30. Johnes PJ (1996) Evaluation and management of the impact of land use change on the nitrogen and phosphorus load delivered to surface waters: The export coefficient modelling approach. J Hydrol 183:323–349
31. Streeter HW, Phelps EB (1925) A study of the pollution and natural purification of the Ohio river. Public Health Bull No. 146
32. Najafabadi MM, Ziaee S, Nikouei A et al (2019) Mathematical programming model for optimization of regional cropping patterns decisions: A case study. Agric Syst 173:218–232
33. Yang G, Li X, Huo L et al (2020) A solving approach for fuzzy multi-objective linear fractional programming and application to an agricultural planting structure optimization problem, vol 141. Chaos, Solitons & Fractals, p 110352
34. Yu H, Liu K, Bai Y et al (2021) The agricultural planting structure adjustment based on water footprint and multi-objective optimisation models in China. J Clean Prod 297:126646
35. Karunakaran KR, Nedumaran S, Chandrasekaran M (2012) Is the current land use pattern in crop agriculture sustainable in the Bhavani Basin of Southern India? Triennial Conference of the International Association of Agricultural Economists
36. Zeng XT, Kang S, Li F et al (2010) Fuzzy multi-objective linear programming applying to crop area planning. Agric Water Manage 98:134–142
37. Tsai YJ, Jones JW, Mishoe JW (1987) Optimizing multiple cropping systems: A systems approach. Trans ASAE 30(6):1554–1561
38. Zheng Z, Gao Y, Nayga RM Jr. et al (2023) Policy reform and farmers' coping strategies: The case of corn price shocks in Heilongjiang Province of China. Rev Dev Econ 27(2):1135–1156
39. Hendricks NP, Smith A, Sumner DA (2014) Crop supply dynamics and the illusion of partial adjustment. Am J Agric Econ 96(5):1469–1491

40. Goetz SJ (1992) A selectivity model of household food marketing behavior in sub-Saharan Africa. Am J Agric Econ 74:444–452
41. Key N, Sadoulet E, de Janvry A (2000) Transaction costs and agricultural household supply response. Am J Agric Econ 82:245–259
42. Lokshin M, Sajaia Z (2004) Maximum likelihood estimation of endogenous switching regression models. Stata J 4:282–289
43. Wooldridge JM (2010) Econometric Analysis of Cross Section and Panel Data, 2nd edn. MIT Press
44. Yin L, You N, Zhang G et al (2020) Optimizing feature selection of individual crop types for improved crop mapping. Remote Sens 12(1):162
45. Hu Q, Wu W, Song Q et al (2015) Recent progresses in research of crop patterns mapping by using remote sensing. Scientia Agricultura Sinica 48(10):1900–1914 (in Chinese)
46. Foerster S, Kaden K, Foerster M et al (2012) Crop type mapping using spectral-temporal profiles and phenological information. Comput Electron Agric 89:30–40
47. Somers B, Asner GP (2013) Multi-temporal hyperspectral mixture analysis and feature selection for invasive species mapping in rainforests. Remote Sens Environ 136:14–27
48. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
49. Huang Q, Tang H, Wu W et al (2013) Remote sensing based dynamic changes analysis of crop distribution pattern: Taking Northeast China as an example. Scientia Agricultura Sinica 46(13):2668–2676 (in Chinese)
50. FAO (1976) A Framework for Land Evaluation. Soils Bulletin No. 32, Food and Agriculture Organization of the United Nations
51. Holzkämper A, Calanca P, Fuhrer J (2013) Identifying climatic limitations to grain maize yield potentials using a suitability evaluation approach. Agric For Meteorol 168:149–159
52. Long TB, Blok V, Poldner K (2017) Business models for maximising the diffusion of technological innovations for climate-smart agriculture. Int Food Agribusiness Manage Rev 20(1):5–23
53. Mnih V, Kavukcuoglu K, Silver D et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
54. Schulman J, Wolski F, Dhariwal P et al (2017) Proximal Policy Optim Algorithms. arXiv:1707.06347.
55. Dong J, Fu Y, Wang J et al (2020) Early-season mapping of winter wheat in China based on Landsat and Sentinel images, 2001–2019. Earth Syst Sci Data 12:3081–3095
56. Liu W, Li S, Cao J et al (2025) Uncovering the spatiotemporal evolution and driving mechanisms of soybean planting area in China from 2000 to 2022. Journal of Integrative Agriculture, in press
57. Shen R, Pan B, Peng Q et al (2023) High-resolution distribution maps of single-season rice in China from 2017 to 2022. Earth Syst Sci Data 15:3203–3222
58. Wolpert DH, Tumer K (2001) Optimal payoff functions for members of collectives. Adv Complex Syst 4(2–3):265–279

59. Foerster J, Farquhar G, Afouras T et al (2018) Counterfactual multi-agent policy gradients. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2974–2982

Figures

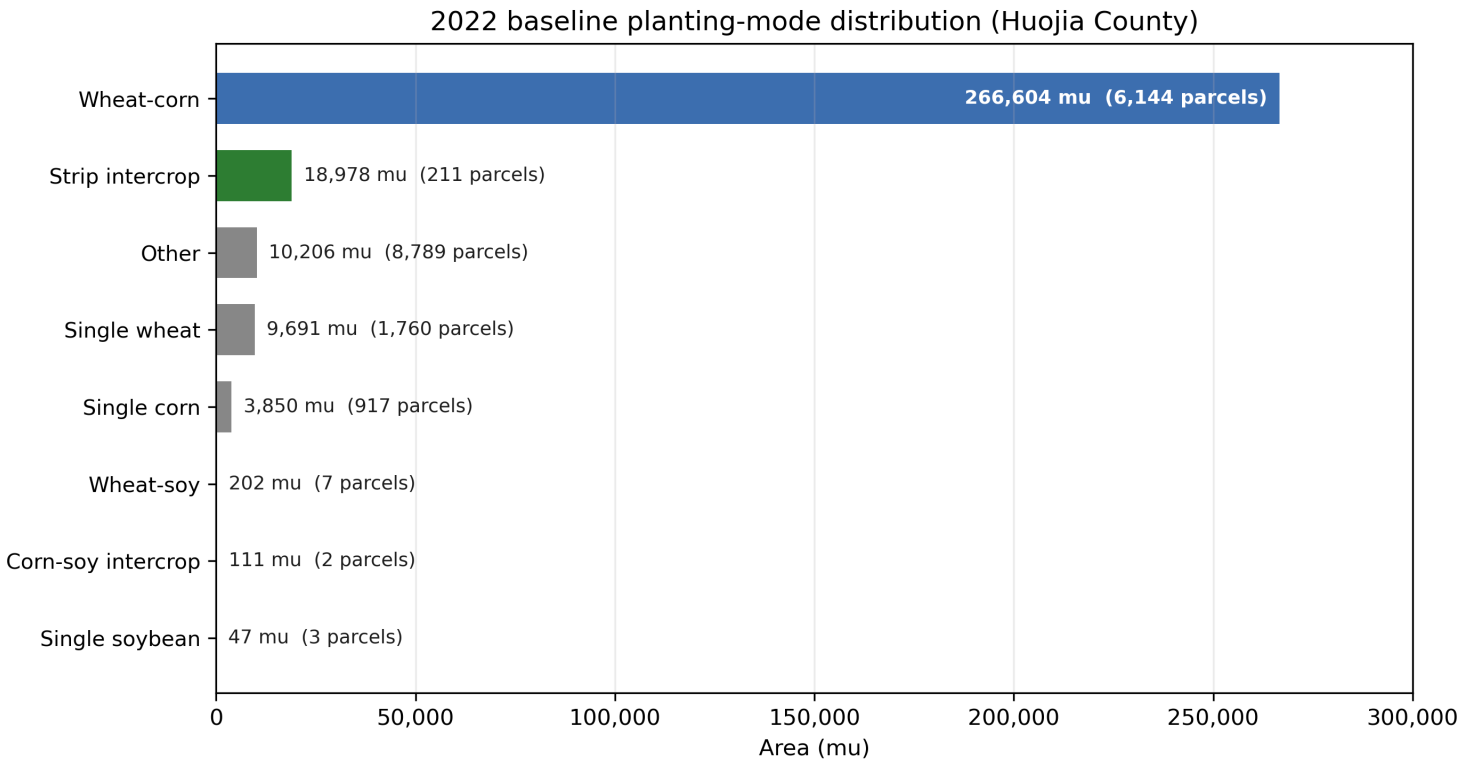


Figure 1

2022 baseline planting-mode distribution. Area and parcel count by planting mode across Huojia County; wheat–corn double cropping (266,604 mu) dominates, with strip-like soybean–corn co-planting (mode 6) at 18,978 mu, an upper-bound practice proxy from the pixel classifier; the enrolled program area in 2022 was about 3,000 mu (Section 2.1).

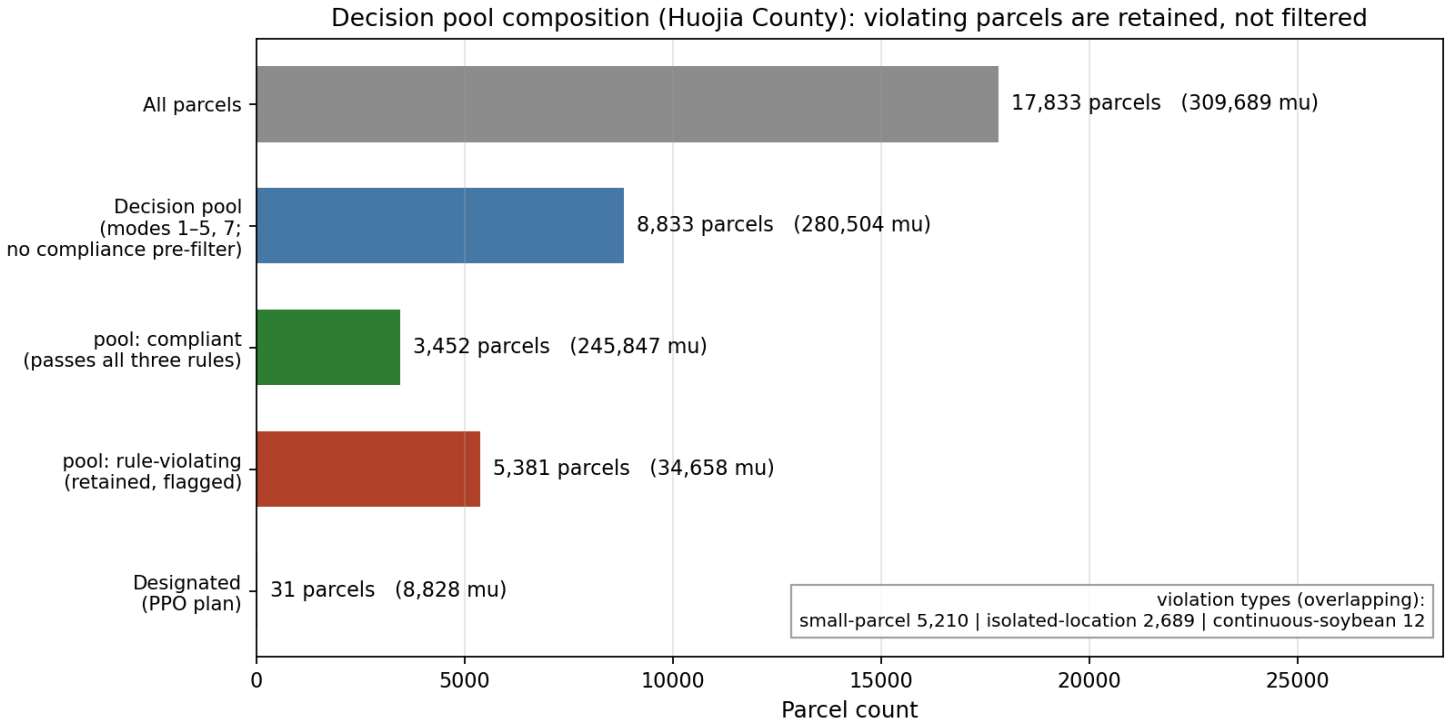


Figure 2

Decision-pool composition. All cultivated non-strip parcels (2022 modes 1–5, 7) enter the pool with no compliance pre-filtering; rule-violating parcels (small-parcel 5,210, isolated-location 2,689, continuous-soybean 12; categories overlap) are retained and flagged. The designated plan (31 parcels, 8,828 mu) is drawn from this mixed pool.

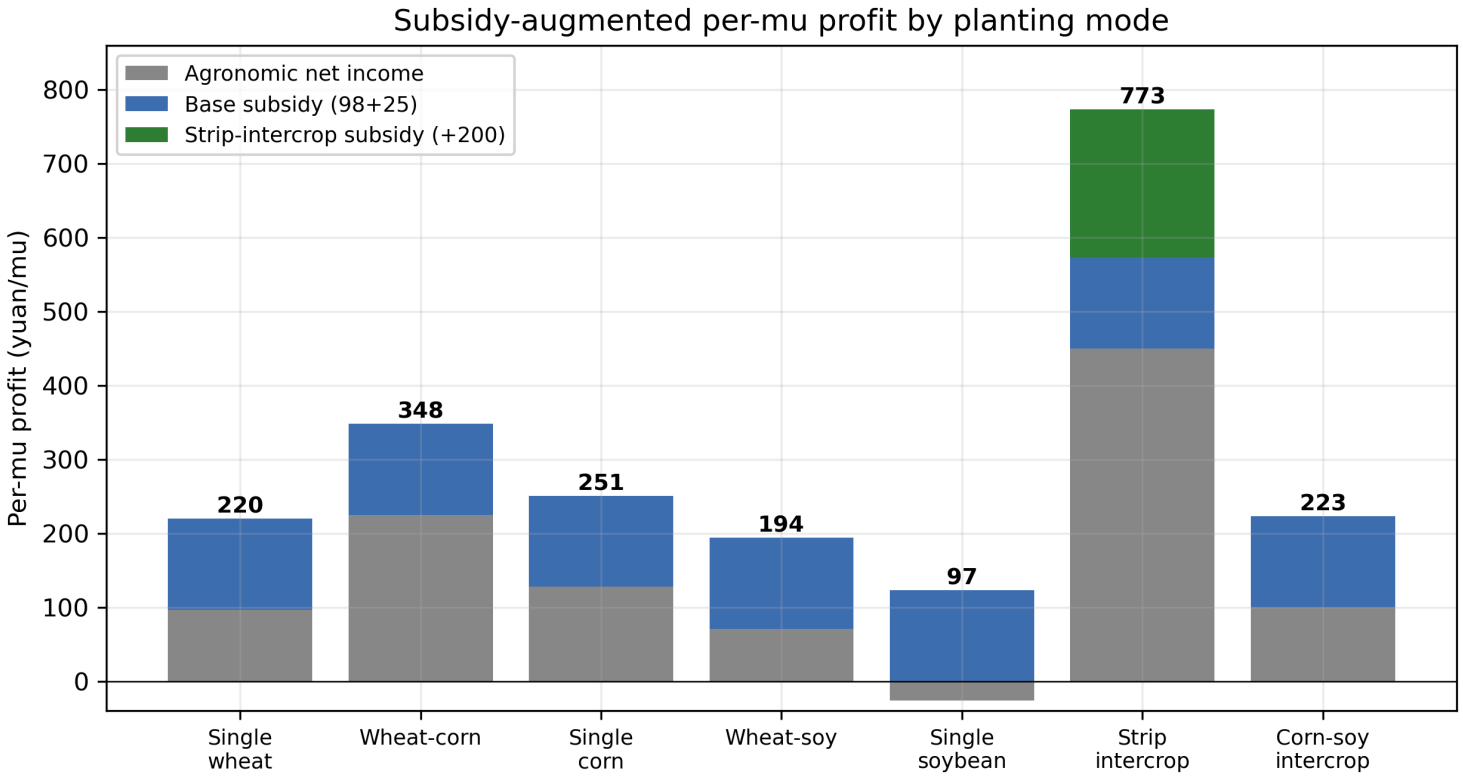


Figure 3

Subsidy structure. Stacked bar chart showing the layered composition of per-mu profit for each of the eight planting modes, separating net income, base subsidy (98 + 25 yuan/mu), and the strip-intercropping specific subsidy (200 yuan/mu).

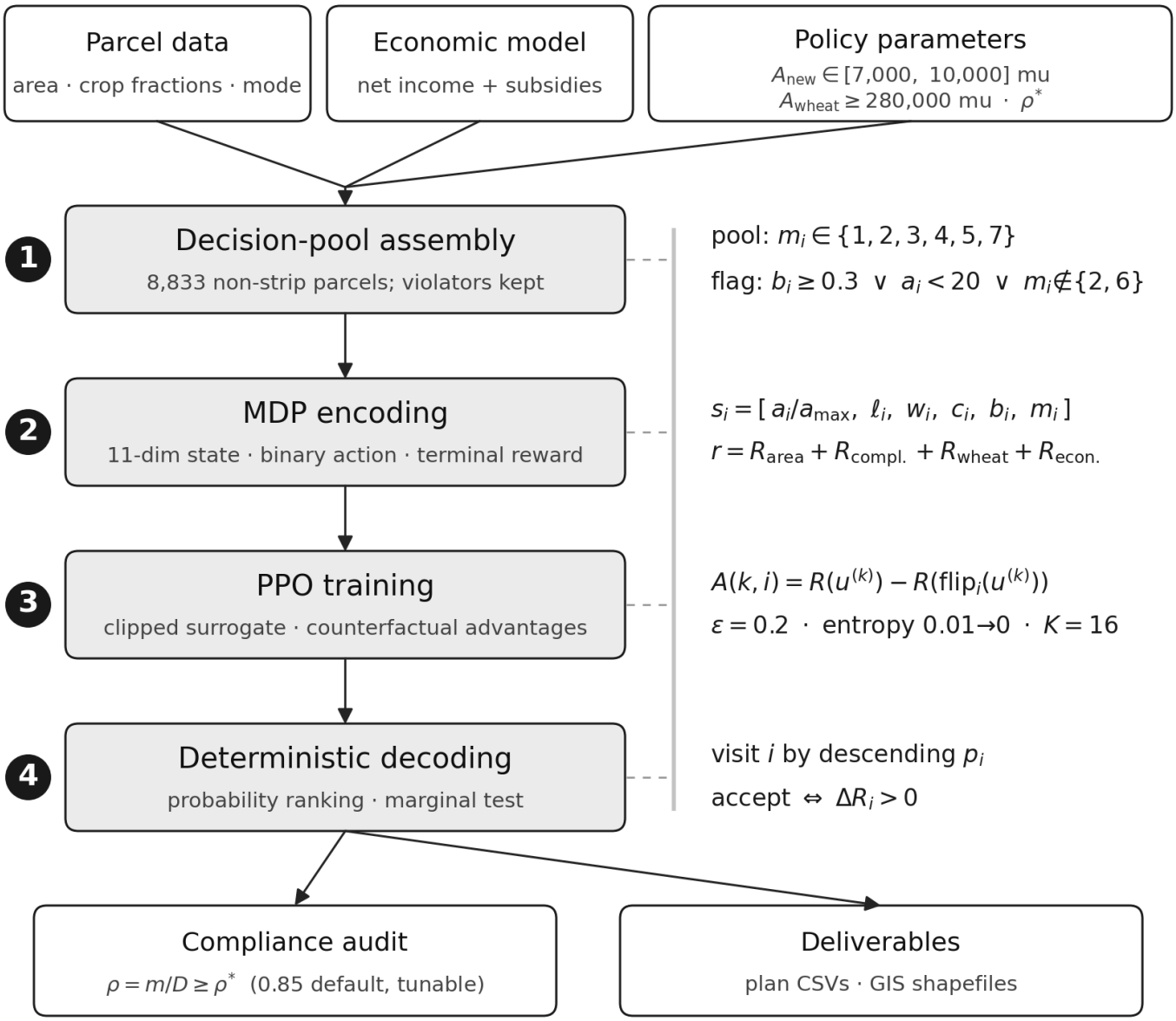


Figure 4

Methodological workflow of the parcel-level PPO framework.

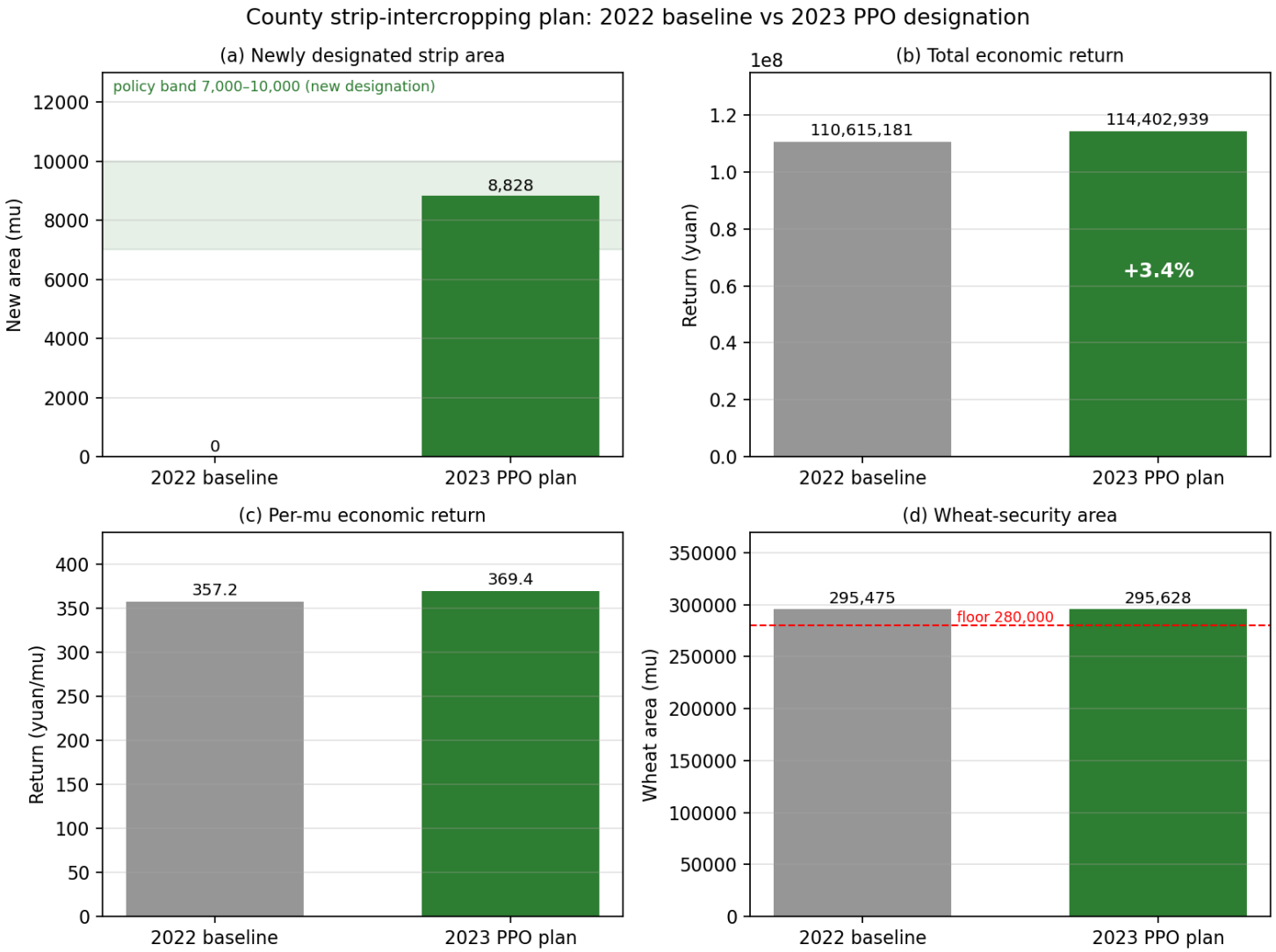


Figure 5

Planning outcome: 2022 baseline versus 2023 PPO designation across four operational metrics: newly designated strip area (with the 7,000–10,000 mu policy band), total economic return, per-mu return, and wheat-security area (with the 280,000 mu floor).

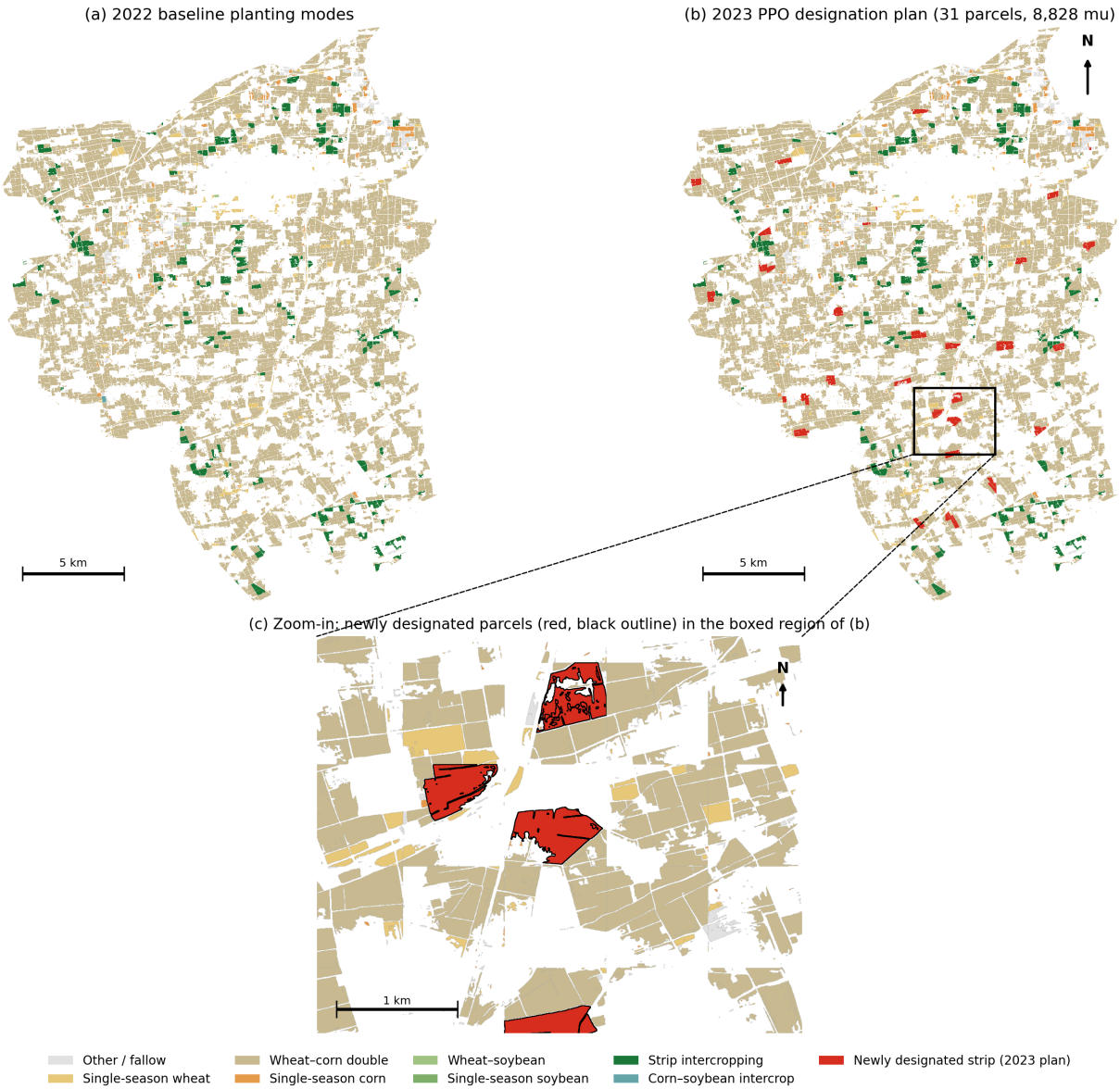


Figure 6

County-scale spatial distribution of planting modes in Huojia County: the 2022 baseline (left) versus the 2023 PPO designation plan (right). Each cultivated parcel is coloured by its planting mode (legend); the 31 parcels newly designated for soybean–corn strip intercropping by the learned policy are highlighted in red. The designated parcels are large wheat–corn parcels distributed across the county; dispersion is expected because this configuration optimizes the contiguity-blind reward of Eq. (2); Section 5.3 shows that adding the contiguity term concentrates designation into fewer clusters.

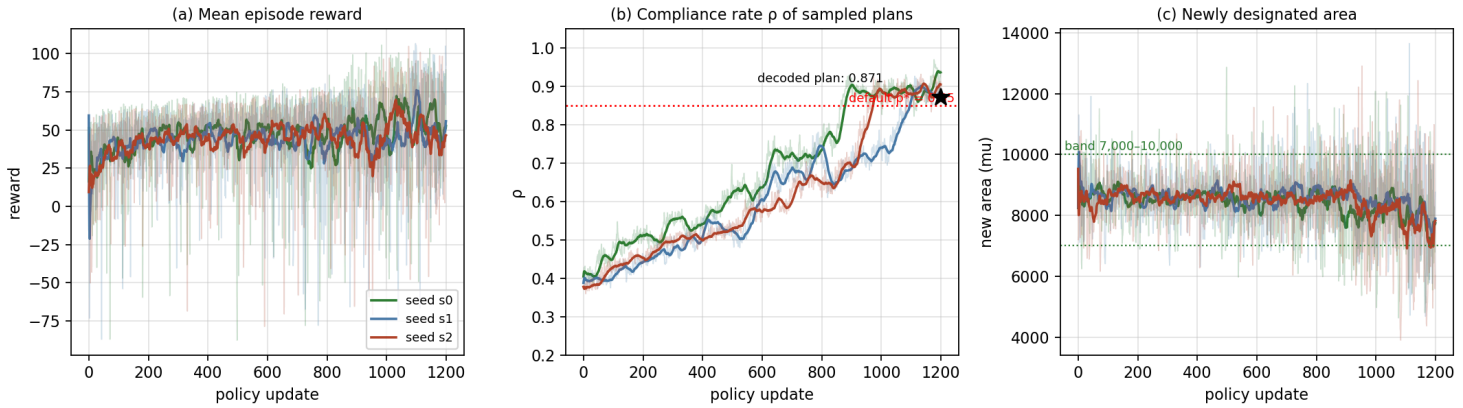


Figure 7

Training dynamics (three seeds). (a) Mean sampled-plan reward; (b) compliance rate ρ of sampled plans rises from the pool composition to a stable plateau at the ρ* = 0.85 floor (star: decoded-plan audit, ρ = 0.871); (c) newly designated area converges into the 7,000–10,000 mu band. Thin lines: raw per-update values; bold: 25-update moving average.

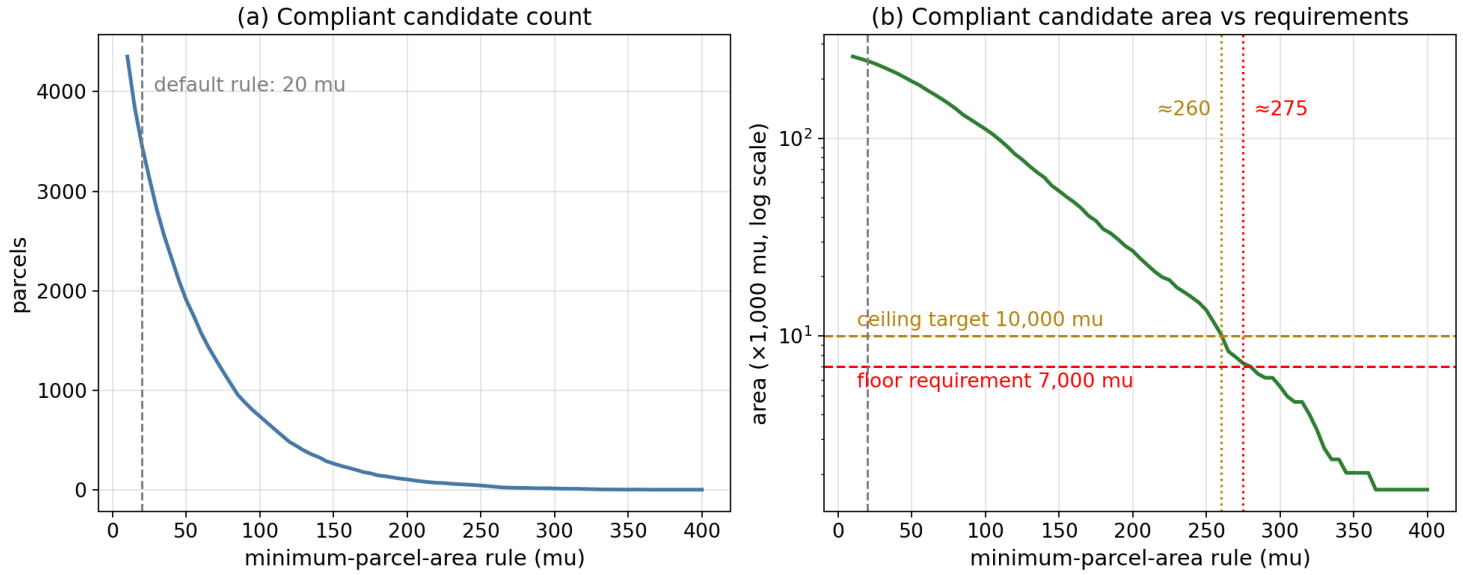


Figure 8

Minimum-parcel-area sensitivity (real parcel data). (a) Compliant candidate count and (b) compliant candidate area as the minimum-area rule sweeps 10–400 mu; horizontal lines mark the new-designation requirements for the band floor (7,000 mu) and the ceiling target (10,000 mu). The ceiling stays attainable up to a rule of about 260 mu and the floor up to about 275 mu (dotted verticals); the grey dashed line marks the default 20 mu rule.

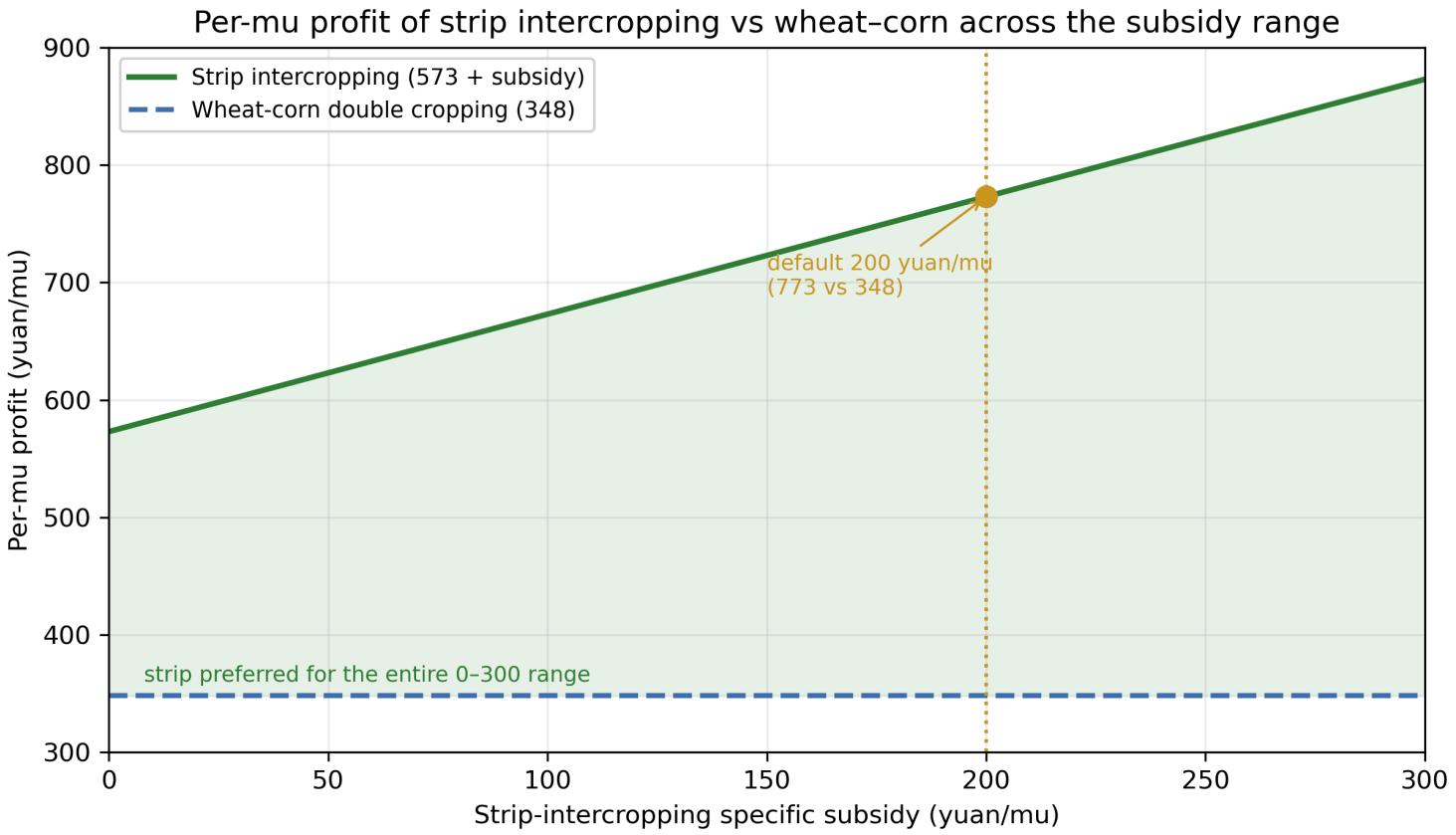


Figure 9

Per-mu profit gradient across the subsidy range (motivating economic model, not a trained-policy result). Per-mu profit of strip intercropping (573 + specific subsidy) versus wheat–corn double cropping (348 yuan/mu) as the strip-specific subsidy varies over 0 to 300 yuan/mu; strip intercropping stays economically preferred across the entire range, with the default 200 yuan/mu yielding 773 versus 348 yuan/mu.

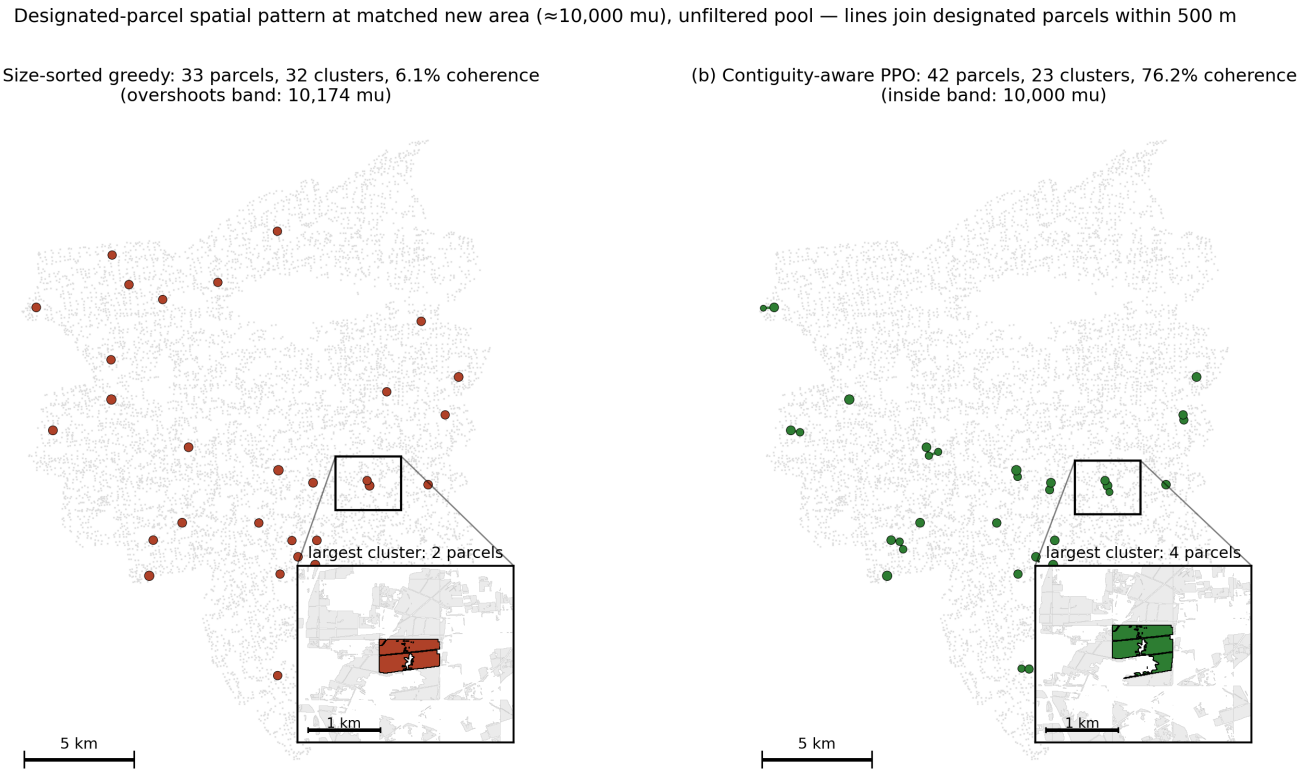


Figure 10

Designated-parcel spatial pattern at matched new area (approximately 10,000 mu) on the unfiltered pool; lines join designated parcels whose centroids lie within 500 m and marker size is proportional to parcel area. (a) The size-sorted greedy heuristic scatters 33 parcels into 32 clusters at 6.1% coherence (a single adjacency edge) and overshoots the band ceiling (10,174 mu); (b) the contiguity-aware PPO policy concentrates 42 parcels into 23 clusters at 76.2% coherence (22 edges), inside the band. Grey dots: pool parcels. The insets show the parcel polygons of each method's largest designated cluster (two parcels for greedy versus four adjacent parcels for PPO) with scale bars on both the county maps and the insets.