Multi-Objective Quantile-Based Reinforcement Learning for Modern Urban Planning

Lukasz Pelcner[1], **Leandro Soriano Marcolino**[1], **Matheus Aparecido do Carmo Alves**[2], **Paula A. Harrison**[3] and **Peter M. Atkinson**[1]
[1]Lancaster University
[2]University of São Paulo
[3]UK Centre for Ecology & Hydrology
{l.pelcner, l.marcolino, pma}@lancaster.ac.uk, mthalves@usp.br and paulaharrison@ceh.ac.uk

Abstract

We present a novel Multi-Agent Reinforcement Learning approach to understand and improve policy development by land-shaping agents, such as governments and institutional bodies. We derive the underlying policy decisions by analyzing the land and developing an intelligent system that proposes optimal land conversion strategies. The aim is an efficient method for allocating residential spaces while considering the dynamic population influx in different regions, jurisdictional constraints, and the intrinsic characteristics of the land. Our main goal is to be sustainable, preserving desirable land types such as forests and fluvial lands while optimizing land organization. We introduce an attractiveness metric that quantifies the proximity to different land types and other factors to optimize land usage. It distinguishes two types of agents: "top-down" agents, which are policymakers and shareholders, and "bottom-up" agents representing individuals or groups with specific housing preferences. Our main objective is to create a synergistic environment where the top-down policy meets the bottom-up preferences to devise a comprehensive land use and conversion strategy. This paper, thus, serves as a pivotal reference point for future urban planning and policy-making processes, contributing to a sustainable and efficient landscape design model.

1 Introduction

Urban planning is a critical domain that requires harmonizing "top-down" policy decisions, implemented by governments and institutional bodies, and "bottom-up" preferences, reflecting the needs and desires of individuals and communities. This duality is particularly significant in addressing the "tragedy of the commons", a scenario where individual incentives clash with the sustainable management of shared resources. Achieving this balance is crucial for the future of human populations, especially in the context of effective use of scarce resources.

This study focuses on solving a specific instance of this duality by introducing a novel machine learning (ML) approach that emphasizes both the theoretical and practical significance of this problem. Specifically, we address residential housing allocation and optimization using a dual agent framework inspired by the work of Bone et al. [2011]. This approach leverages a robust experimental setting to demonstrate how intelligent systems can navigate trade-offs between policy-driven objectives and individual preferences.

We propose a comprehensive framework that incorporates two distinct types of agents: "top-down" agents, such as policymakers and institutional stakeholders, and "bottom-up" agents, representing individuals or groups with specific housing preferences. By employing this dual agent framework, we aim to design methods tailored to the unique challenges of this class of problems, focusing on preserving ecological balance while optimizing land use. Therefore, we introduce:

(i) **Quantile-Optimized Land Use (QOLU) Algorithm for Top-Down Agents**, which employs deep reinforcement learning (RL) to model strategic land use planning. QOLU agents optimize multiple goals, such as minimizing agricultural land conversion, preserving proximity to freshwater sources, and ensuring that new developments increase spatial proximity with existing urban and suburban areas. QOLU agents aim to act as stewards of the environment, safeguarding the continuity of woodlands, agricultural expanses, and other pivotal land use types.

(ii) **Neural Network-based Bottom-Up Investor Agent (BUIA) Algorithm**, a decentralized planning model which uses limited observability to prioritize land use changes based on historical data and local preferences. This agent leverages neural networks to identify profitable and sustainable opportunities within a 2 km radius.

Our framework addresses the technical gaps in balancing top-down and bottom-up approaches, while also offering practical innovations by explicitly optimizing multi-objective trade-offs present in our context. This novel integration allows for a synergistic strategy that adapts to diverse geographical, climatic, and socio-economic conditions.

The remainder of this paper details the problem formulation, methodology, and a comparative analysis against established benchmarks from the literature. We also present experimental results, validating our approach and discussing its broader implications for future urban planning, along with potential extensions to enhance scalability and applicability.

2 Related Work

Balancing economic development with environmental stewardship is a key theme in computational urban planning. Early approaches often employed methods such as neural-network-based cellular automata [Li and Yeh, 2002] or probabilistic graphical models [Bone *et al.*, 2011] to simulate land-use changes under diverse constraints. Multi-agent systems were later introduced to model interactions among heterogeneous stakeholders, as in studies of distributed decision-making for traffic coordination [Wiering, 2000] and sustainable zoning [Zheng *et al.*, 2023]. In parallel, single-agent reinforcement learning (RL) techniques leveraged spatial information to predict urban expansion, but often prioritized a single global objective [Qian *et al.*, 2023].

Recent advances in distributional RL have offered more robust solutions for tasks involving uncertainty and conflicting goals by estimating a return distribution instead of a point estimate [Bellemare *et al.*, 2017]. Methods such as implicit quantile networks [Dabney *et al.*, 2018] allow finer control over multi-objective trade-offs, making them appealing for land management scenarios where ecological interests can clash with development demands. Work in multi-objective RL provides a broader framework for learning policies that balance competing criteria, surveyed extensively in Roijers et al. [2013] and Liu et al. [2015], while specific applications to urban growth reveal the viability of RL-based algorithms for complex spatial environments [Stetter *et al.*, 2024].

Despite these developments, bridging top-down regulations (e.g., environmental protection) with bottom-up stakeholder preferences (e.g., local housing markets) remains challenging, though early attempts at "modeling-in-the-middle" [Bone *et al.*, 2011] underscored the need for linking policy instruments to agent-level decisions. We address this gap by integrating a quantile-based multi-objective RL framework, suitable for safeguarding critical land types, with decentralized agents that capture localized incentives. This hybrid design seeks to produce more context-aware and environmentally aligned outcomes than purely top-down or purely bottom-up models.

3 Methodology

We frame our land management problem as a modified multi-agent Partially Observable Stochastic Game (POSG), aiming to combine macro-level policy objectives with micro-level stakeholder preferences. The methodology centers on two distinct agent types: *top-down* Quantile-Optimized Land Use (**QOLU**) agents that adopt distributional RL to guard critical land types, and *bottom-up* Neural Network-based Investor Agents (**BUIA**) that capture localized incentives based on an *attractiveness metric*.

3.1 Problem Formulation

We formulate our problem as a POSG due to its capability to model the uncertainties and agents in our context – mimicking the realistic constraints in urban planning. Our formulation tries not only to capture the inherent complexity of urban environments but also allows for robust policy evaluation against conflicting objectives. Our POSG is given by the tuple:

$$(\mathbf{\Phi}, \mathbf{S}, \mathbf{A}, T, R, \mathbf{Z}, O),$$

where $\mathbf{\Phi} = (\phi_1, \dots, \phi_n)$ is the set of agents, partitioned into top-down shareholders ($\mathbf{\Phi}_{TD}$) and bottom-up investors ($\mathbf{\Phi}_{BU}$). Each state $s \in \mathbf{S}$ describes a set of $1\ \mathrm{km}^2$ parcels with unique attributes (e.g., current land usage, geographic constraints). The action space $\mathbf{A} = \mathbf{A}_1 \times \cdots \times \mathbf{A}_n$ encodes land-conversion decisions, and T is the transition function specifying the probability of moving from one configuration of parcels to another given a joint action. Each agent observes only partial information, defined by an observation function $O\colon \mathbf{S} \times \mathbf{A} \times \mathbf{Z} \to [0, 1]$. The reward function:

$$R(s, \mathbf{a}) \;=\; \sum_{i=1}^{n} w_i \, R_{\phi_i}(s, a) \;+\; R_{\text{society}}(s, \mathbf{a})$$

integrates heterogeneous agent objectives via the weights w_i. Contrary to the usual POSG formulation, although each agent computes its own objective R_{ϕ_i}, the environment supplies *one shared team reward* $R(s, a)$, encouraging fully cooperative learning. For a top-down agent, R_{ϕ_i} may include terms for preserving woodlands and agricultural land and maintaining proximity to freshwater, whereas the societal reward $R_{\text{society}}(s, \mathbf{a})$ captures the net public benefit of protecting overall forest and agricultural areas (parameterized by α and β). Bottom-up agents receive rewards linked to attractiveness values of newly developed parcels, reflecting immediate local gains.

3.2 Attractiveness Metric

A key component for bottom-up decisions is the *attractiveness metric* that assigns to each parcel a score based on proximity to natural features (e.g., forests, fluvial areas) and existing urban centers. Let $\mathcal{B}_{\text{attractiveness}}$ store the scores A_i for each parcel i, with higher values signifying increased desirability for residential or commercial development. The probability of an investor agent targeting a parcel is proportional to its attractiveness, thereby balancing land conversion pressures against ecological and jurisdictional constraints. Post-conversion, the attractiveness scores update to reflect changes in neighboring parcels, allowing the system to evolve iteratively.

3.3 Quantile-Optimized Land Use (QOLU)

Top-down agents adopt a quantile-based distributional RL approach designed to preserve environmental continuity while accommodating necessary development. Each top-down agent or *stakeholder* may optimize multiple objectives, from minimizing agricultural land loss to maintaining sufficient distance from fragile ecosystems. The RL algorithm produces a unified *quality metric* indicating how favorable the conversion of a non-residential parcel is to the agent's weighted goals.

QOLU Algorithm. We instantiate QOLU with a deep RL model that applies quantile regression to learn a distribution over returns, rather than a single expected value. This formulation is initialized by parameters $\boldsymbol{\psi}^0$ and updated iteratively:

$$\boldsymbol{\psi}^{t+1} \;=\; \boldsymbol{\psi}^{t} - \eta \, \nabla L(\boldsymbol{\psi}^{t}),$$

where η is the learning rate. The loss $L(\boldsymbol{\psi}^t)$ is computed over transitions (s, a, r, s') using the quantile regression function

$$L(\phi) \;=\; \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \rho_{\tau_j}\big(y_{ij} - Q_{\boldsymbol{\psi}}(s_i, a_i, \tau_j)\big),$$

Algorithm 1 Quantile-Optimized Land Use (QOLU) Algo.

```
Initial policy parameters $\theta$, empty replay buffer $D$, exploration schedule $Epsilon$, number of atoms $N$, $V_{min}$,$V_{max}$, $\gamma$ discount factor
for each actor $k$ running in parallel do
    Initialize the actor's environment state $s$
    for each step $t$ do
        Select action $a$ by exploiting noisy network parameters $\theta$ or exploratory action based on $Epsilon$
        Execute action $a$ in the environment to get reward $r$ and new state $s'$
        Store the transition $(s, a, r, s')$ in $D$
        Update $s \leftarrow s'$
    end for
end for
Sample a minibatch of transitions $(s, a, r, s')$ from $D$
for each transition in minibatch do
    Calculate n-step return
    $R_t = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \max_{a'} Q(s_{t+n}, a'; \theta)$
    Update target distribution $Z = R_t$ for the corresponding atom
end for
For Double Q-learning, use $\arg\max_a Q(s', a; \theta)$ to select an action and $\theta^-$ to evaluate it, resulting in $Z = R_t + \gamma Z(s', \arg\max_a Q(s', a; \theta); \theta^-)$
With a dueling architecture, separate the value and advantage streams in the network, then combine them for the final Q values calculated as:
$Q(s, a; \theta) = V(s; \theta) + \left(A(s, a; \theta) - \sum_{a'} \frac{A(s,a';\theta)}{|A|}\right)$
Perform a gradient descent step on the Kullback-Leibler divergence $D_{KL}(Z||Z(s, a; \theta))$ with respect to the network parameters $\theta$
Every $t_n$ steps reset $\theta^- = \theta$
```

with $\tau_j \in (0, 1)$ specifying the quantile index and ρ_{τ_j} the quantile loss function. For each sampled transition, we update the distribution of future returns, thereby capturing the range of possible outcomes under uncertain land conversion scenarios. Pseudocode for QOLU closely follows a Distributional DQN with quantile regression, as shown in Algorithm 1.

Once trained, each QOLU agent assigns a quality score to candidate conversions. Summing these scores across all stakeholders provides a consensus-driven measure that balances ecological, economic, and policy-related objectives. Table 1 presents the details about the architecture and hyper-parameters used for the QOLU implementation.

Atom support. Following Bellemare *et al.* [2017], we approximate the return distribution by a categorical ("atom-based") distribution supported on a fixed, finite interval.

$$[\,V_{\min},\, V_{\max}\,] \subset \mathbb{R}.$$

The interval is divided into N equally-spaced atoms

$$z_i = V_{\min} + i\Delta z, \quad \Delta z = \frac{V_{\max} - V_{\min}}{N - 1}, \qquad 0 \le i < N,$$

which play the role of "canonical" returns.

Algorithm 2 Bottom-Up Investor Agent (BUIA) Algo.

```
Initialize model parameters $\theta$
Define feature set $\mathbf{X}$ capturing local land-use data
for each decision point do
    Compute logits $\mathbf{Z} = \mathrm{NeuralNetwork}(\mathbf{X}; \theta)$
    Convert $\mathbf{Z}$ to distribution $\mathbf{P} = \mathrm{softmax}(\mathbf{Z})$
    Sample an action $a$ from $\mathrm{Categorical}(\mathbf{P})$
    Execute $a$ (e.g., convert selected parcels)
    Observe reward $R_{\mathrm{NN}}$ based on change in attractiveness
end for
```

3.4 Neural Network-Based Bottom-Up Investor Agent (BUIA)

While top-down QOLU agents promote global objectives, *bottom-up* agents capture the micro-level incentives of individuals or groups seeking property development. Each BUIA operates under limited observability, constrained to a small radius of nearby parcels. By exploiting local attractiveness scores and historical land-use changes, BUIA agents can identify economically and ecologically favorable parcels to convert.

BUIA Algorithm. Each BUIA agent uses a neural network that processes local features, such as surrounding land types and updated attractiveness values. The network outputs a probability distribution over potential development actions, typically selecting parcels that maximize expected profitability while aligning with partial ecological constraints. At each decision step, the agent:

1. Gathers local context (e.g., land cover, neighbors' attractiveness updates).
2. Feeds these features into a neural model (parameters θ).
3. Obtains a categorical action distribution via softmax.
4. Samples an action a and executes the corresponding land-use change.

The reward $R_{\mathrm{NN}}(s, a) = \mathrm{Attract}(s') - \mathrm{Attract}(s)$ encourages conversions that increase desirability over time. Algorithm 2 defines BUIA's pseudo code, showing the critical role of the *attractiveness metric* in guiding bottom-up decisions.

Once a BUIA completes its localized conversion, the corresponding land records update, and newly computed attractiveness scores propagate to both bottom-up and top-down agents. This interplay establishes a feedback loop in which

Conv. Block	3×3 kernels, $32 \rightarrow 64 \rightarrow 128$ filters
Flatten Layer	
Residual MLP	2 fully-connected layers (256 units) + skip connection around the pair
Quantile Head	51 atoms with $V \in [-200, 200]$
Optimizer	Adam
Mini-batch size	256
Discount factor γ	0.99
Exploration	ϵ-greedy (schedule not fixed)

Table 1: QOLU's architecture and hyper-parameters information.

Conv. block	1 convolutional layer, kernel and filter counts matching the input
MLP	$128 \rightarrow 128 \rightarrow 64$
Output	Softmax over current candidate parcels

Table 2: BUIA's architecture information.

micro-level parcel changes feed into macro-level planning objectives, creating an evolving land-use landscape that balances individual development goals and broader policy constraints. Table 2 presents the details for BUIA's implementation.

3.5 Environment Representation and Iteration

The experimentation within the simulated environment adheres to a sequential decision-making process, wherein the actions proposed by various algorithms are collated before any updates to the environmental state are enacted. This section delineates the procedural environment workflow, emphasizing the role of decision-making in optimizing land usage.

Sequential Decision Making. In the simulated environment, decision-making is executed in a sequential manner. The iterative process unfolds as follows:

1. Each algorithm proposes its actions based on the current state of the environment.
2. The proposed changes to the map are applied only after all decisions have been made, ensuring a synchronized update across the entire environment.

This approach ensures that each algorithm operates with the same information, and changes are made based on an aggregation of all decisions at the end of each iteration.

Map Division and QOLU Agent Allocation. The environment is partitioned into blocks of 40×40 pixel units, referred to as parcels, as shown in Figure 1. These parcels represent the jurisdiction of individual QOLU agents, who act as stakeholders with vested interests in the land use outcomes. Each parcel is assigned an attractiveness metric, reflecting the desirability of the neighborhood characteristics within that segment.

$$\mathcal{B}_{\text{attractiveness}} = \{A_1, A_2, \ldots, A_n\} \tag{1}$$

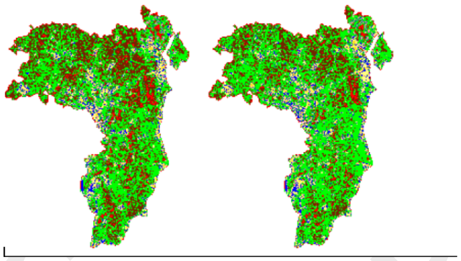


Figure 1: Southeast UK region, 2015, before (left map) and after 1000 iterations of QOLU (right map).

In Equation 1, $\mathcal{B}$ denotes a buffer storing attractiveness scores for every parcel, A_i denotes the attractiveness rating of parcel i, and n is the total number of parcels.

Decision Buffer and Sampling. A decision buffer is constructed to hold the parcels from which they are sampled. The probability of a parcel being sampled is directly proportional to its attractiveness rating:

$$P(\text{sampled} \mid A_i) \propto A_i \tag{2}$$

In Equation 2, the weighted sampling ensures that parcels with higher attractiveness are more likely to undergo decision-making processes by the agents, thus ensuring faster convergence from a machine learning perspective, and higher average satisfaction with the change.

Neural Network Agent Decision Making. After sampling, the BUIA agent selects positions within the non-urban land that will be converted into urban land types. This agent's decisions are informed by the updated land attractiveness metrics, influencing the development pattern of the urban landscape.

Reward Allocation and Iteration Completion. Once the BUIA agent has made its selections, the new parcel data is fed back to the Stakeholder QOLU agents. Each QOLU agent then allocates rewards based on the degree of compliance with their individual objectives:

$$R_{\text{QOLU},i} = f(\ parcel\_data, \text{objectives}_i\ ),$$

where f represents the reward function specific to each QOLU agent i, and their pre-determined objectives for each agent i.

Simulation Iteration and Land Conversion. An iteration of the simulation is deemed complete once the reward allocation is finalized. It is noteworthy that during each iteration, 50 parcels are sampled with replacement, allowing for the potential conversion of up to 11 land types into urban areas.

Iterative Process and Convergence. The simulation proceeds iteratively, with the described sequence of steps repeating. Convergence towards an optimized state is measured by the stabilization of attractiveness metrics and reward distributions across successive iterations.

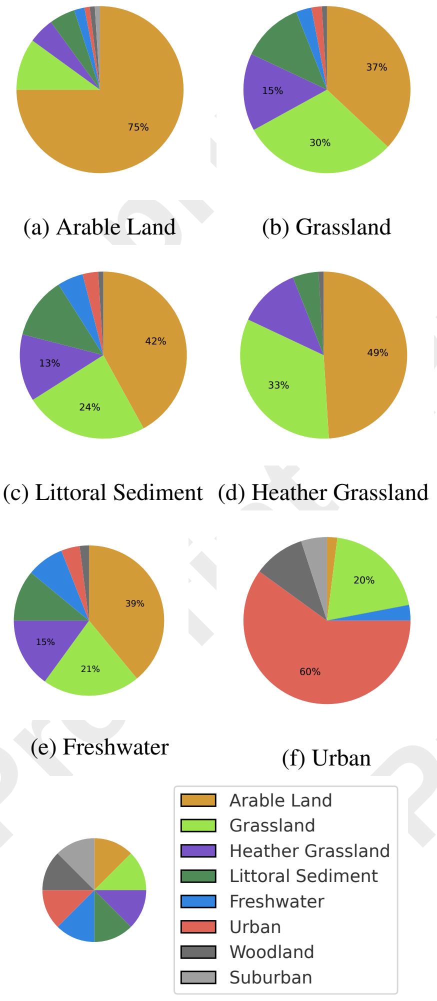
This experimental framework is designed to provide insights into the effectiveness of the proposed algorithms in managing land development in a way that balances individual objectives with broader societal and environmental considerations.

3.6 Metrics

Land Preference Metric. The land preference metric is defined based on the probability distribution of land types $\mathbf{L}$ for a given cell. Let $P(\mathbf{L})$ be the distribution. The land preference metric, denoted as **LP**, is calculated as:

$$\mathbf{LP} = \sum_{i=1}^{n} w_i \cdot P_{\phi_{TD_i}}(\mathbf{L}) + P_{society}(\mathbf{L}) \tag{3}$$

In Equation 3, w_i represents the weight for each top-down agent, $P_{\phi_{TD_i}}$ is the land preference function for the i-th high policy agent, and P_{society} is the societal land preference function taken from a dataset and visualized in Figure 2.



(a) Arable Land (b) Grassland

(c) Littoral Sediment (d) Heather Grassland

(e) Freshwater (f) Urban

(g) Colours Legend

Figure 2: Land preferences represent natural co-occurrence of land types in the real world, guiding future balance preservation.

Well-being Metric. The well-being metric considers the housing capacity and unoccupied houses. Let H_{total} be the total housing capacity and H_{occupied} be the occupied housing. The well-being metric is calculated as $\mathbf{WB} = \frac{H_{\text{occupied}}}{H_{\text{total}}}$, assessing how effectively the housing capacity is utilized.

4 Experimental Settings

Overview. The evaluation process involves:

(i) Initialization: Each agent is initialized with a set of objectives, which can include minimizing the loss of woodland and arable land, maximizing the distance of developed parcels from bog, mountain, and heath areas, and minimizing the distance to urban areas.

(ii) Simulation: The agents get sampled on a 50×50 grid (each cell represents a land parcel). The bottom-up agents propose changes to the land parcels based on their objectives, and top-down agents appraise the solution.

(iii) Reward Calculation: The agents assign rewards to each change based on how well the change aligns with their objectives. Rewards are calculated using a combination of individual rewards (based on specific objectives) and a global reward (based on the overall impact on the environment and urban planning).

(iv) Iteration: The process is repeated for 1000 iterations, with each iteration representing a new set of proposed changes and evaluations.

(v) Aggregation: The results from all iterations are aggregated to provide a comprehensive assessment of the agent's performance. This includes calculating the average rewards and analyzing the distribution of land use changes over the entire grid.

The primary metrics used for evaluation include:

(i) Proximity to Desired Land Types: Measuring how close the developed parcels are to bog, mountain, and heath areas.

(ii) Preservation of Land Types: Assessing the extent to which woodland and arable land types are preserved.

(iii) Urban Connectivity: Evaluating how well the developed parcels are connected to existing urban areas.

Stakeholders Preferences and Scenarios. To evaluate the performance of our proposed method, we performed experiments considering three stakeholders' preferences: (i) maximizing the distance of developed parcels to specific land uses (bog, mountain, heath) aggregated into a single land type; (ii) minimizing the loss of woodland and arable land types due to development, and; (iii) minimizing the distance of newly developed parcels to urban or suburban lands.

From these preferences, we created four scenarios, each emphasizing a different preference. The scenarios and their respective weights are outlined in Table 3.

Conversion Metric. A critical metric introduced in this study is the land type distribution alignment C_{align}:

$$C_{\text{align}}(s) = \frac{\sum_{i=1}^{N} \left(\text{commonality}(s_i) - \text{commonality}(\hat{s}_i)\right)^2}{N}$$

where N is the number of parcels, s_i is the set of land types neighboring parcels i before agent actions, and $\hat{s}_i$ is the set after agent actions. The function commonality(s) measures the frequency of the most common neighboring land type to parcel i. This metric assesses the agents' capacity to preserve existing land type distributions that are assumed to reflect historical human land-shaping preferences as shown in Figure 2.

	Land Use	**Woodland**	**Urban**
Scenario 1	0.33	0.33	0.33
Scenario 2	0.50	0.25	0.25
Scenario 3	0.25	0.50	0.25
Scenario 4	0.25	0.25	0.50

Table 3: Preference settings for each different scenario.

Baselines. We propose three baselines from the state-of-the-art for our experiments:

1. the Modelling-in-the-Middle (MitM)'s approach, proposed by Bone et al. [2011].
2. the LToS algorithm, proposed by Yi et al. [2021].
3. the P-MADDPG, proposed by Pelcner et al. [2024].

5 Results

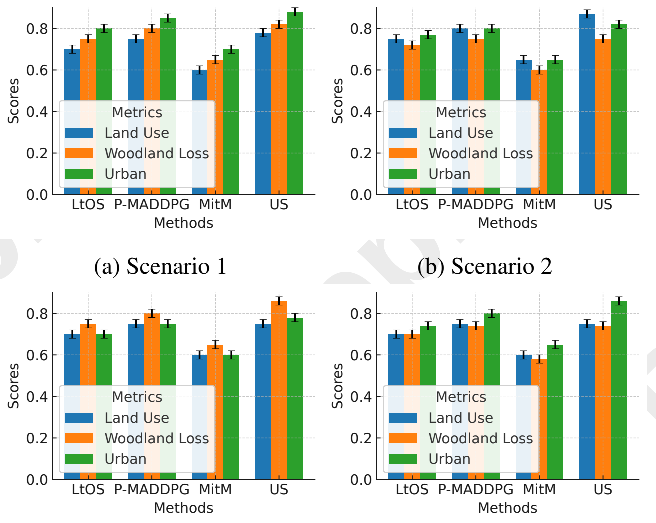
In the result analysis, we conducted paired *t*-tests for each scenario and metric to evaluate the confidence in our results. All scores are presented normalized with standard errors.

Scenario 1: Uniform Weights. Under the uniform weights scenario (Table 4), our proposed method significantly outperforms the baseline methods in all three metrics (*t*-test analysis with $p < 0.05$). This demonstrates the robustness of our approach when all stakeholders are given equal importance.

Scenario 2: High Emphasis on Distance to Land. In Table 5, we can see that our method achieves the highest score (0.85) for this metric, significantly outperforming LToS (0.75) and P-MADDPG (0.80). However, for Woodland Loss and Distance to Urban, our method performs similarly to P-MADDPG (0.75 and 0.80, respectively). This suggests that while our approach is highly effective in prioritizing Distance to Land Use, it maintains competitive performance in other metrics.

Scenario 3: High Emphasis on Woodland Loss. Table 6 shows that our method achieves the highest score (0.85) for this metric among baselines. While the performance on Distance to Land Use is slightly lower (0.75), it is still comparable to P-MADDPG (0.75). The *t*-tests indicate that the improvements in Woodland Loss are statistically significant ($p < 0.05$), demonstrating our method's capability to effectively reduce the impact on woodland areas.

Scenario 4: High Emphasis on Distance to Urban. Table 7 shows the results for this setting. Our method achieves the highest score (0.85) for this metric. Although the performance on Distance to Land Use and Woodland Loss (0.75) is not significantly better than for P-MADDPG, the overall performance indicates that our method effectively prioritizes urban proximity while maintaining balance across other metrics.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 3: Performance comparison under different scenarios.

Summary. While our method demonstrates significant improvements across most scenarios, the mixed results in certain cases can be attributed to the inherent trade-offs between the different metrics. For instance, optimizing Distance to Land Use might slightly compromise Woodland Loss and *vice-versa*. These trade-offs highlight the complexity of urban planning and the need for flexible and adaptive approaches. Our method's ability to perform well across various scenarios, significantly improving key metrics, affirms its robustness and effectiveness in addressing diverse urban planning challenges.

6 UKCEH Dataset

In environmental research, the availability of extensive and long-term datasets is crucial for the development and application of advanced algorithms. The UK Centre for Ecology & Hydrology (UKCEH) stands as a key contributor, offering a valuable repository of data that not only informs scientific endeavors but also facilitates practical applications in real-life scenarios. This section highlights this dataset's utility as a

Method	Land	Woodland	Urban
LToS	0.70 ± 0.01	0.75 ± 0.02	0.80 ± 0.03
P-MADDPG	0.75 ± 0.02	0.8 ± 0.03	0.85 ± 0.02
MitM	0.60 ± 0.02	0.65 ± 0.02	0.70 ± 0.01
US	0.8 ± 0.03	0.85 ± 0.02	0.9 ± 0.01

Table 4: General performance under Scenario 1.

Method	Land Use	Woodland	Urban
LToS	0.75 ± 0.02	0.70 ± 0.02	0.75 ± 0.02
P-MADDPG	0.80 ± 0.01	0.75 ± 0.02	0.80 ± 0.02
MitM	0.65 ± 0.02	0.61 ± 0.01	0.65 ± 0.02
US	0.85 ± 0.02	0.75 ± 0.02	0.80 ± 0.01

Table 5: Distance to Land Use's performance under Scenario 1.

Method	Land Use	Woodland	Urban
LToS	0.70 ± 0.01	0.75 ± 0.02	0.70 ± 0.02
P-MADDPG	0.75 ± 0.02	0.80 ± 0.02	0.75 ± 0.02
MitM	0.60 ± 0.02	0.65 ± 0.02	0.60 ± 0.01
US	0.75 ± 0.02	0.85 ± 0.02	0.75 ± 0.02

Table 6: Woodland Loss's performance under Scenario 2.

Method	Land Use	Woodland	Urban
LToS	0.70 ± 0.01	0.70 ± 0.01	0.75 ± 0.02
P-MADDPG	0.75 ± 0.02	0.75 ± 0.02	0.80 ± 0.02
MitM	0.60 ± 0.02	0.60 ± 0.01	0.65 ± 0.02
US	0.75 ± 0.02	0.75 ± 0.02	0.85 ± 0.02

Table 7: Distance to Urban's performance under Scenario 3.

crucial piece in developing and applying an RL MAS.

As users of the UKCEH data, our primary objective is to leverage the land use dataset to enhance our understanding of the environmental dynamics in the southwest region of the United Kingdom. Spanning the years 2015 to 2021, this dataset serves as a vital component in our larger effort to create and implement a reinforcement learning multi-agent system, designed to navigate the complexities of real-world scenarios.

Our interest lies in the practical application of this dataset as we work towards developing algorithms that can adapt and learn within dynamic environmental contexts. By incorporating the UKCEH's land use data, we aim to enrich our understanding of the region, enabling our reinforcement learning multi-agent system to operate effectively in real-life settings.

7 Discussion & Conclusions

Discussion. The duality problem in urban planning, characterized by the interaction between "top-down" policies and "bottom-up" preferences, poses significant challenges in achieving sustainable resource management. This study addresses these challenges by proposing a dual-agent framework, combining the QOLU for policymakers and BUIA for individual stakeholders. This discussion highlights the implications, limitations, and potential extensions of our approach.

Relevance to the "Tragedy of the Commons". At the core of our research is the concept of the "tragedy of the commons", where unregulated individual actions can deplete shared resources. By explicitly modeling the duality between centralized decision-making and decentralized preferences, our framework systematically addresses resource depletion. The integration of QOLU and BUIA enables a synergistic optimization of societal goals, such as minimizing the loss of ecologically significant land types, while addressing the localized preferences of individuals seeking residential housing.

Strengths and Novel Contributions. Our framework contributes to the field of urban planning by addressing several critical aspects: **(i) Multi-Objective Optimization:** The QOLU algorithm demonstrates its ability to balance competing objectives, such as preserving agricultural and woodland areas while ensuring urban connectivity. **(ii) Adaptability:** The BUIA algorithm leverages local data and individual preferences, offering a decentralized approach to land-use planning that complements the global strategies of QOLU agents. **(iii) Transferability:** While our study focuses on urban residential planning, the proposed framework can be extended to other domains, such as agricultural land management, ecosystem conservation, and flood risk mitigation.

Conclusions. We developed a multi-agent system for land-use optimization, introducing two novel algorithms: Quantile-Optimized Land Use (QOLU) and the neural network-based Bottom-Up Investor Agent (BUIA). QOLU agents balance competing land-use policy objectives, while BUIA agents select the most attractive parcels for development. Our approach models uncertainty by capturing full return distributions and balancing ecological, economic, and societal goals within a POSG framework. These objectives are adaptable, allowing incorporation of additional metrics and domain-specific knowledge for broader applicability. We benchmarked our framework against three state-of-the-art baselines across four scenarios reflecting diverse stakeholder preferences. Our method consistently outperformed these baselines, achieving statistically significant improvements in key metrics such as land-use alignment, woodland preservation, and urban proximity. QOLU agents effectively minimized ecologically critical land loss, ensured appropriate land separations, and enhanced urban planning. BUIA agents prioritized high-attractiveness parcels, enabling strategic, desirable development. This work advances sustainable land management by offering a robust framework that combines RL algorithms and neural network-based decision-making. Future extensions could incorporate additional environmental and socio-economic factors to further improve adaptability and impact.

Limitations and Challenges. Despite its strengths, the proposed framework has several limitations that warrant further exploration: **(i) Scalability:** While our experiments demonstrate efficacy on a grid of 1 km^2 land units, scaling the framework to larger regions with higher agent densities may introduce computational challenges. **(ii) Data Dependency:** The accuracy of the BUIA agent depends heavily on the availability and quality of local land-use data. In regions with sparse data, the performance of the framework may be affected. **(iii) Dynamic Factors:** Our current implementation assumes relatively static socio-economic and environmental conditions. Incorporating dynamic factors, such as population growth and climate change, remains an area for future work.

Future Direction. Building on our results and insights, avenues for future research are proposed: **(i) Incorporation of Dynamic Models:** Introducing dynamic population and environmental models can enhance the realism and applicability of the framework. **(ii) Improved Scalability:** Leveraging distributed computing and advanced optimization techniques can address scalability challenges in larger geographical settings. **(iii) Integration of Additional Metrics:** Expanding the framework to include socio-economic factors, such as income distribution and housing affordability, would make the framework more comprehensive. **(iv) Real-World Validation:** Collaborating with urban planners and policymakers to validate the framework in real scenarios could bridge the gap between theoretical research and practical implementation.

Although our current implementation is tailored to a specific domain (UKCEH dataset) [UK Centre for Ecology & Hydrology, 2023], our framework is naturally extendable. We are exploring transfer learning techniques and parameter tuning to adapt to different geographic regions with varying regulatory and socio-economic conditions. However, the promising results presented here already demonstrate the system's ability to handle complex, real-world dynamics.

Acknowledgments

This paper was partially funded by the New Approaches to Data Science Programme, Data Science of the Natural Environment (grant EP/R01860X/1) and São Paulo Research Foundation (FAPESP) grant no. 2024/15649-7. We thank Marta Maslankowska for reviewing the early draft of this work.

References

[Bellemare *et al.*, 2017] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[Bone *et al.*, 2011] Christopher Bone, Suzana Dragicevic, and Roger White. Modeling-in-the-middle: bridging the gap between agent-based modeling and multi-objective decision-making for land use change. *International Journal of Geographical Information Science*, 25(5):717–737, 2011.

[Dabney *et al.*, 2018] Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1104–1113. PMLR, 2018.

[Li and Yeh, 2002] Xia Li and Anthony Gar-On Yeh. Neural-network-based cellular automata for simulating multiple land use changes using gis. *International Journal of Geographical Information Science*, 16(4):323–343, 2002.

[Liu *et al.*, 2015] Chunming Liu, Xin Xu, and Dewen Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2015.

[Pelcner *et al.*, 2024] Lukasz Pelcner, Matheus Aparecido do Carmo Alves, Leandro Soriano Marcolino, Paula Harrison, and Peter Atkinson. Incentive-based marl approach for commons dilemmas in property-based environments. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 2414–2416, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems.

[Qian *et al.*, 2023] Kejiang Qian, Lingjun Mao, Xin Liang, Yimin Ding, Jin Gao, Xinran Wei, Ziyi Guo, and Jiajie Li. Ai agent as urban planner: Steering stakeholder dynamics in urban planning via consensus-based multi-agent reinforcement learning, 2023.

[Roijers *et al.*, 2013] Diederik M. Roijers, Peter Vamplew, Richard Dazeley, and Gerhard Weiss. A survey of multi-objective sequential decision-making. In *Journal of Artificial Intelligence Research*, volume 48, pages 67–113. AI Access Foundation, 2013.

[Stetter *et al.*, 2024] Christian Stetter, Robert Huber, and Robert Finger. Agricultural land use modeling and climate change adaptation: A reinforcement learning approach. *Applied Economic Perspectives and Policy*, 46, 05 2024.

[UK Centre for Ecology & Hydrology, 2023] UK Centre for Ecology & Hydrology. Ukceh land cover map (2015–2021 series). https://www.ceh.ac.uk/services/land-cover-map, 2023. Accessed: 2025-06-10.

[Wiering, 2000] Marco Wiering. Multi-agent reinforcement learning for traffic light control. In *Proc. ICML 2000*, pages 1151–1158, 2000.

[Yi *et al.*, 2021] Yuxuan Yi, Ge Li, Yaowei Wang, and Zongqing Lu. Learning to share in multi-agent reinforcement learning. *arXiv preprint arXiv:2112.08702*, page 16, 2021.

[Zheng *et al.*, 2023] Yu Zheng, Hongyuan Su, Jingtao Ding, Depeng Jin, and Yong Li. Road planning for slums via deep reinforcement learning, 2023.