Reinforcement Learning to Enhance Geospatial Analysis of Clandestine Graves in Mexico

Valeria Vera Lagos[1]

[1] Georgetown University

ABSTRACT

In this work, a Reinforcement Learning (RL) method for a geospatial analysis of clandestine graves in Mexico is presented. For law enforcement authorities and families looking for missing people, the issue of clandestine graves and the discovery of remains is a complicated and sensitive one. The goal of the study is to assess area reduction exploration and visualize common trajectories to find clandestine graves using RL. The work uses various techniques to increase the effectiveness of the RL model, including estimating state values, action values, and optimal policies and adding function approximators.

Keywords: Reinforcement Learning, Geospatial.

1 Introduction

The issue of clandestine graves in Mexico is one that needs to be addressed right away and is extremely concerning. In the context of the drug war and human trafficking, there are currently more than 100,000 people missing, and more than 2,000 illegal graves have been found in 75% of the states. The emotional toll and social effects of these disappearances on Mexican society are highlighted by the thousands of parents and family members who have organized collectives throughout the country to assist one another in the search for the remains of their missing children. The founder of “The voice of the Missing” group, Mara Luisa Nez, said that having a family member go missing is a "Total pause in your life”.

Clandestine graves are defined as locations where one or more people were illegally and/or anonymously taken into custody. The existence of these graves, and the discovery of remains within them, can provide critical information for law enforcement agencies and families searching for those who have gone missing. However, finding and excavating these graves is a risky task that frequently necessitates cooperation from different groups. This task considers the traits of victims of missing persons in a specific place and time and the geographical circumstances of the areas where the events occurred, as well as the societal, political, economic, and cultural contexts of the population.

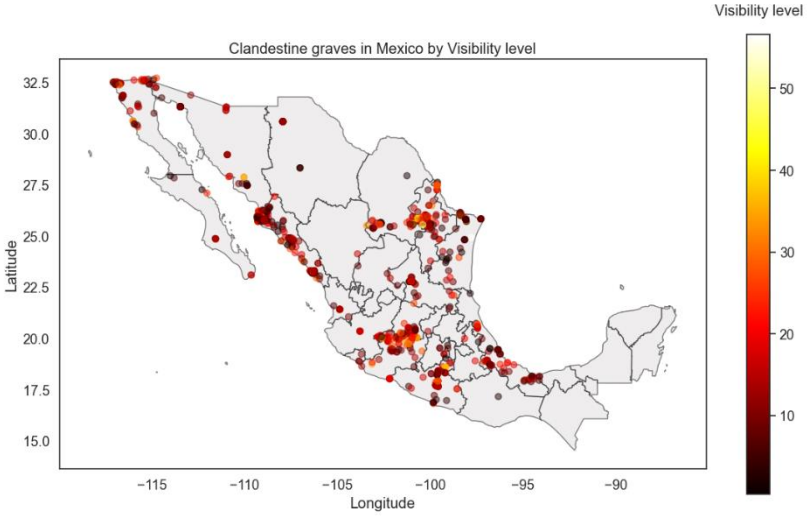

Figure 1

The objective of this work is a Geospatial Analysis of clandestine graves in the states Veracruz and Sinaloa using Reinforcement Learning to evaluate exploration techniques for area reduction and visualize common trajectories, considering the seriousness and complexity of the issue.

2 Methods

2.1 Data

Data was collected by using the “Plataforma de Transparencia” from the INAI institution in Mexico. Each of the states federal agencies delivered the information of clandestine graves location in a printed pdf that was later transformed into a csv. To obtain the visibility level and distance to the closes city the portal delivered in [2] was used as a search engine.

2.2 Environment definition

The following terms will be applied to define the components of the problem of Clandestine Graves that are relevant to the Reinforcement Learning (RL) model.

- Agent: The decision-maker that engages with the environment is the RL agent.
- Environment: The agent is an external system that interacts with the environment. In this problem, two states, Sinaloa and Veracruz, will serve as models. For the agent to explore, the two states will be divided into a "grid" environment.
- State: Environment's current condition is reflected in its state. It includes every information required for the agent to make a decision. The state is fully observable in this work.
- Action: The agent's decision at each time step is the action. It influences the environment's condition and decides how much the agent will be rewarded. Two sets of movements—up, down, right, and left—by one space (4 actions) and by two spaces each (24 actions) are defined in this work.
- Reward: An agent receives a scalar feedback signal as compensation for an action by the environment. Its function is to direct the agent toward its objective. The aim of RL typically consists of maximizing the expected cumulative reward over a series of time steps. The reward for this work will be determined by taking into account geographical factors that are similar among clandestine graves such as visibility range and distance to the closest city. To determine a reward for each grid space, the following formula was used:

  $$\text{reward} = \text{normalize}(((1/\text{visibility} + \text{distance})\,100)\text{graves})$$

  Since both visibility and distance decrease significantly when a grave is present, their summation was inverted and multiplied by a 100 so that the graves' characteristics influence how much reward is maximized. Finally, this value was multiplied by the number of graves and normalized.

- Policy: It is a mapping between states and actions. It describes the agent's behaviour in relation to its environment. The objective in this work is to identify a policy that

maximizes the overall reward. The policy will be used to interpret the locations more likely to contain graves. A variety of policy evaluation techniques will be used to find optimal policies.

- e-greedy policy picks an action at random with e probability; and with probability, it selects the action that maximizes the action-value function. This way, we continue to explore all state-action pairs while selecting the best actions we identify with high likelihood. [1]

- Terminal state: Even though the objective of the work is not for the agent to arrive to a terminal state but rather explore the spaces with higher rewards, a terminal state is defined as the state with the highest reward.

- Function approximation: Function approximation allow to generalize from seen states to unseen states and save space and model the dynamics of a process from which we have observed data. In this work a Feed-Forward Neural Network is utilized as function approximation.

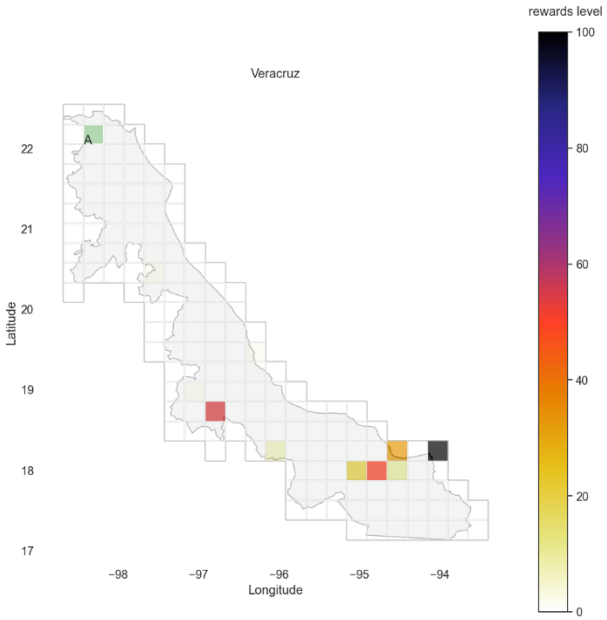

Figure 2

Figure 2 shows the full definition of the grid environment with rewards level for spaces with clandestine graves.

2.3 Algorithms

According to [1] in Reinforcement Learning (RL), our goal is to obtain (near) optimal policies. To accomplish this, different methods such as the estimation of state values, action-values and optimal policy are used, which can be combined to enhance the efficiency of a RL model. The main distinction between this method is the quantity being estimated such as the policy, the action-value or the state-value function. The following section describes them thoroughly.

Approximating state-value function involves estimating the expected long-term reward of being in a particular state. The goal is to find the value of each state, which can be used to determine the best action to take in each of them. A common method for approximating the state-value function is the First-visit Monte Carlo which is part of a collection of Monte Carlo (MC) estimation methods that (book) describes as "make estimations through repeated random sampling". First-visit Monte Carlo estimates state and action values using sample trajectories of complete episodes.

Furthermore, estimating action-values require to approximate the expected long-term reward of taking a particular action in a particular state to find the best one. One common algorithm for estimating action-values is the SARSA (on-policy) algorithm which uses a temporal difference learning approach to update the estimate of the action-value function. It learns by experiencing the environment and updating the state-action value at every time step. The agent samples the next step using the policy it is learning. Meaning that in order to update state-action values, SARSA therefore considers the control policy that the agent is moving under.

Finally, to estimate the optimal policy involves looking for the best possible policy that an agent can follow in each environment to maximize its reward. Meaning to determine the optimal action to take in each state to achieve the highest possible long-term reward. Multiple methods are performed to estimate the optimal policy:

- Monte Carlo ES (Exploring Starts). Starting the trajectory with a random action chosen in a random initial state and then applying the policy as usual. This guarantees that every state-action pair is selected at least once, allowing us to calculate the action-values. The disadvantage of this strategy is to repeatedly initialize episodes at random.

- On-policy first-visit Monte Carlo control algorithm (for e-soft policies) A state might be visited multiple times during the episode. The episode may be referred to as a first visit if the state is visited for the first time. As a result, the first-visit MC method averages the returns following state's first visits to estimate the value of that state.

- Off-policy Monte Carlo control algorithm. Off-policy methods evaluate or enhance a policy that differs from the one that was used to generate the data, "book" mentions Off-policy methods, when combined with function approximators, could have issues with converging to a good policy. They follow the behaviour policy while learning and improving the estimation policy.

- SARSA (on-policy TD control). This is a method for estimating the optimal policy in which the agent follows a given policy and uses the SARSA algorithm as described before. The difference relies on having a soft policy, such as e-greedy, to continuously try every action for a given state to improve the policy based on the estimated action-values.

- Q-learning (off-policy TD control). Similar to SARSA but also an off-policy method. Based on the observed rewards and the action with the highest estimated value in the following state, Q-learning updates the estimates of the action-values. The policy is being updated to be greedy with regard to the most recent estimates.

3 Results and discussion

The algorithms described in the previous section have been put to evaluation through multiple experiments. The time it took for the algorithm to run was generally impacted by the quantity of states and rewards. Although most of the algorithms took a long time to explore the space without reaching the terminal states or occasionally getting locked on a "good" local maximum, techniques like First-visit Monte Carlo were showing how the reward was increasing exponentially while visiting more and more states. However, given that a KNN could predict with 93% accuracy and an FFNN with 95% accuracy, we could understand why using function approximation (FA) was able to resolve the problem of local maxima and shorten the time it took to reach the terminal state.

Trajectory without FA on Sinaloa

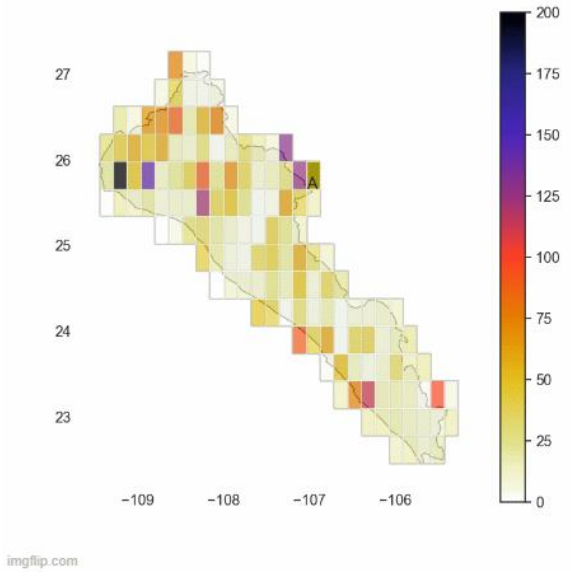

Figure 3

Trajectory with FA in Veracruz

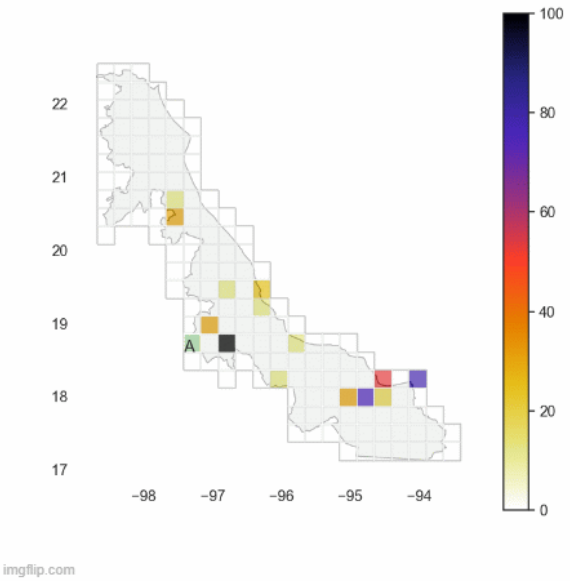

Figure 4

Given the definition of our problem three metrics were assessed using a SARSA on-policy algorithm with and without function approximation (FA). The multiple lines on the graphs represent different "runs" of the algorithm. The first is Rewards, which represents the total of rewards earned throughout the exploration. The number of steps represents how many times the agent left its previous location, and distance represents the sum of the distances between each step in the episode.

Given the vast number of states with rewards, Figure 5 illustrates how the rewards typically rise exponentially.

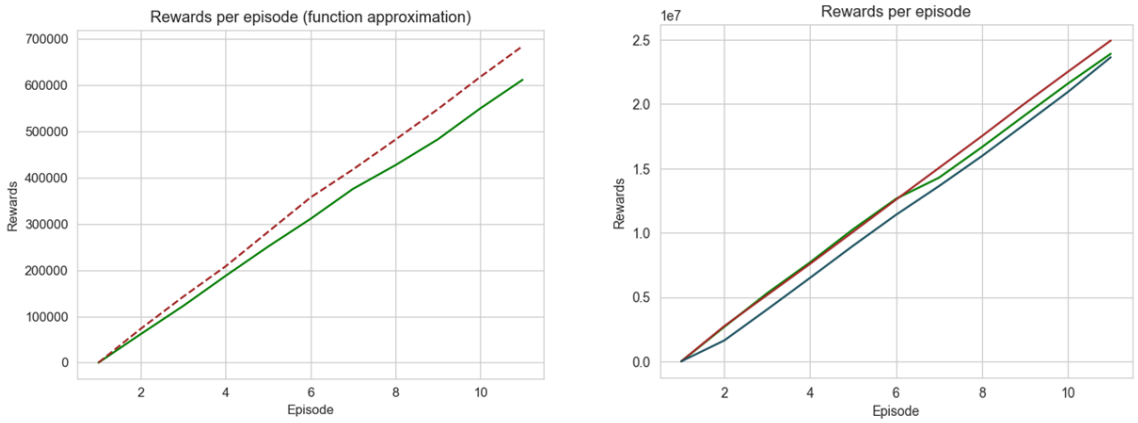

Figure 5

Even though the non-FA method yields a large number of rewards, from the agent's trajectory is observed that it is related to the local maxima problem because the agent keeps returning to the same state.

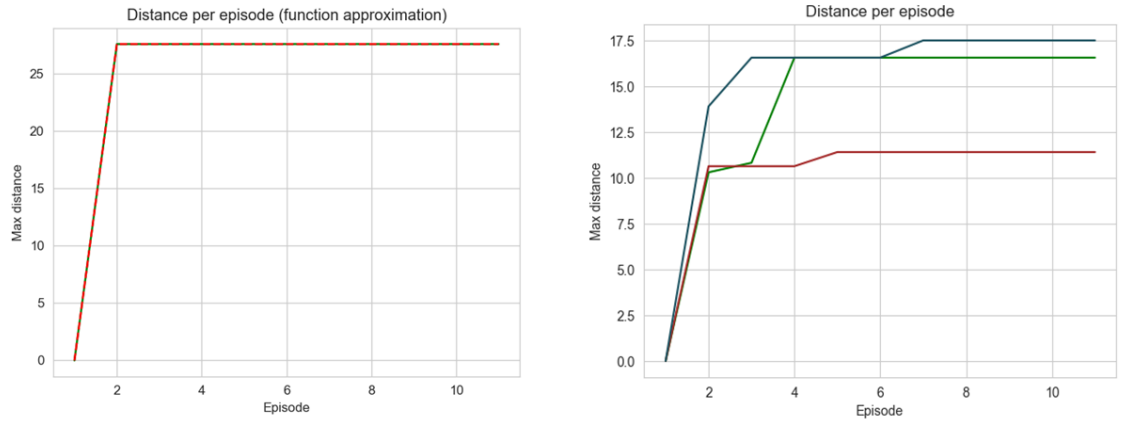

Figure 6

Figure 6 illustrates this issue, with distances demonstrating how the FA agent consistently explores greater distances than the non-FA agent.

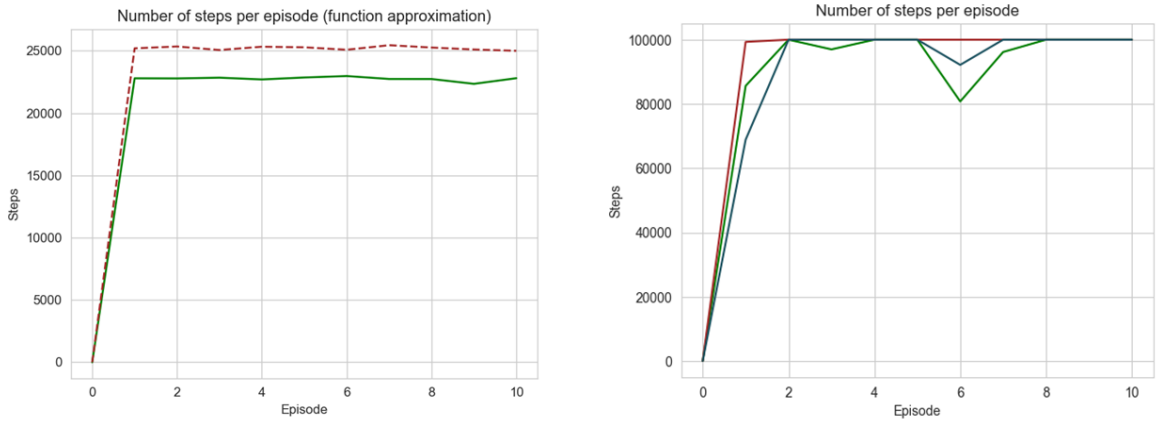

Figure 7

Finally, Figure 7 shows the non-FA agent has 4 times more steps per episode, which is related to the actions for each model since The FA agent can move in a larger area with fewer movements.

Based on these findings, we are able to observe the effects of using accurate models as function approximations, particularly the effects of using a deep learning model in conjunction with a reinforcement learning environment.

4 Conclusions

In conclusion, there is a serious and complicated issue with hidden graves in Mexico that demands attention. Unfortunately, the usage of Reinforcement Learning to assess exploration techniques for area reduction would not be the ideal environment for this problem given that the agent is learning a fixed space with generated rewards. However, this work showed he complexity of defining a Reinforcement Learning environment when geographic factors are considered, as well as the consequences of using function approximations like neural networks. Finally, the work provides an alternative approach for modelling various solutions to a challenging but crucial issue.

References

[1] Mastering Reinforcement Learning with Python. (2020, December 1). O'Reilly Online Learning. https://www.oreilly.com/library/view/mastering-reinforcement-learning/9781838644147/

[2] Silván-Cárdenas, José Luis, Ana Josseline Alegre Mondragon, and Karime González Zuccolotto. "Potential distribution of clandestine graves in Guerrero using geospatial analysis and modelling." iGISc. 2019.