Precision Agriculture Optimization based on Multi-Armed Bandits Algorithm: Wheat Yield Optimization under Different Temperature and Precipitation Conditions

Qikang Huang[1]

Hainan International College, Communication University of China, Hainan, China

Abstract. Climate change and the growing unpredictability of environmental elements such as temperature and precipitation present considerable challenges to contemporary agriculture. Data-driven algorithms present promising solutions by offering more precise tools for optimizing crop yields and resource efficiency to tackle these challenges. Among these approaches, the multi-armed bandit (MAB) algorithm effectively balances exploration and exploitation, showcasing considerable potential for optimizing agricultural decision-making. This study investigates four widely utilized Multi-Armed Bandits (MAB) algorithms: Explore Then Commit (ETC), Upper Confidence Bound (UCB), Asymptotically Optimal UCB, and Thompson Sampling (TS). The objective is to optimize wheat yield under varying temperature and precipitation conditions while also assessing the effectiveness of these different algorithms in achieving this goal. The experiment demonstrates that the UCB algorithm is optimal for analyzing data on total precipitation during the growth of wheat. . Furthermore, the TS algorithm significantly outperforms others in analyzing flat temperature data throughout the wheat growth period. Therefore, the Asymptotically Optimal UCB algorithm can identify the most suitable rainfall conditions for wheat growth in a changing environment. In contrast, the TS algorithm can determine the optimal temperature requirements for wheat growth under similar environmental fluctuations. These insights assist agricultural practitioners in timely adjusting their strategies to enhance crop yield. Additionally, it provides a model for those who want to use the MAB algorithm to improve agricultural yields.

1. Introduction

Climate change has already begun to exacerbate the effects of agricultural production[1], with global warming driven by greenhouse gas emissions being a significant contributing factor. Given the increasing complexity and unpredictability of environmental factors[2] such as temperature and precipitation[3][4], optimizing agricultural production has become

1Corresponding author:202229013059n@mails.cuc.edu.cn

a significant challenge for contemporary agriculture[5]. The advent of data-driven algorithms presents an up-and-coming solution to these challenges, offering farmers more precise tools to enhance crop yields and resource efficiency. The multi-armed bandits (MAB) algorithm balances exploration and exploitation and demonstrates significant potential in domains such as disease detection[6] and the scientific management of agricultural products[7]. It is not difficult to perceive that these algorithms can assist in optimizing decisions based on various uncertainties by constantly learning from the environment[8].

In 2020, Lattimore, T., and Szepesvari, C. analyzed the application of the MAB algorithm in various fields, particularly the balance between exploration and utilization in the natural environment[9]. Through the testing of UCB and TS algorithms, the performance of these algorithms in various application scenarios is verified. The article presents an exhaustive analysis of multiple MAB algorithms, yet its emphasis is more on technical implementation than empirical studies in specific agricultural contexts. Reinforcement learning and multi-armed bandits algorithms are likely employed in the agricultural field, and these technologies can effectively address the issues related to crop management and decision-making[10]. In 2024, D Baudry, R Gautron, and others refined the classical MAB algorithm to assist agricultural practitioners in adjusting nitrogen fertilizer strategies for enhancing yields. In 2022, Daniel Dooyum Uyeh, Senorpe Hiablie, and others employed the Thompson Sampling algorithm to attain the optimal location of sensors in agriculture[11]. Christian Stetter and Robert Huber employ machine learning algorithms for modeling agricultural and forestry land use in the context of climate change.[12]However, existing studies have not utilized the MAB algorithm in specific critical environments related to agriculture, such as typical temperature and precipitation conditions. These factors frequently exert a decisive influence on the production of agricultural products. If the MAB algorithm is utilized as the optimal strategy for continuously adjusting and enhancing production in a complex and dynamic environment, it will significantly benefit agricultural practitioners.

The study employs several extensively utilized MAB algorithms to optimize wheat yield under diverse environmental conditions. Specifically,The study pays attention to two crucial variables: the average temperature and the total rainfall throughout the wheat growing season.These factors are well-known to exert a considerable influence on wheat yield, rendering them optimal targets for agricultural optimization via intelligent algorithms.

The primary objective of our research is to conduct a comparative analysis of the performance among four prominent MAB algorithms: Explore Then Commit (ETC), Upper Confidence Bound (UCB) [13], Asymptotically Optimal UCB, and Thompson Sampling (TS) [14]. The trial is designed to assess how these algorithms can reduce cumulative regret when making decisions related to temperature and rainfall conditions to optimize wheat yields. Cumulative regret is the difference between the return of the chosen action and that of the optimal action. Furthermore, algorithms with lower cumulative regret demonstrate a more effective convergence towards the best decision, making this metric particularly relevant for evaluating performance in this context.

The study utilize a dataset sourced from Kaggle, specifically the Agriculture Crop Yield dataset. This dataset encompasses data on six distinct crops, including wheat, as well as various environmental factors. The study employs this information to identify the most effective algorithms for agricultural decision-making. The findings not only provide valuable insights into the MAB algorithm's potential for optimizing crop production, but also contribute significantly to the broader field of precision agriculture. It also can customize different MAB algorithms to optimize wheat yield according to different

environmental conditions. It can adapt to environmental changes in the climate change to achieve the purpose of precision agriculture.

2. Experimental Design

2.1 Data Source

The dataset employed in this experiment is sourced from Kaggle and comprises 1 million samples. The dataset comprises six distinct crops: wheat, rice, corn, and others. It also provides detailed information on the regions where these crops are cultivated and environmental variables such as rainfall and temperature. Furthermore, it includes the yields associated with the most significant varying factors. In this experiment, only wheat species are selected as the research subjects. The two factors that exert the most significant influence on yield are identified as variables: the average temperature during the crop growing season (measured in degrees Celsius) and the indirect rainfall received throughout this period (measured in millimetres).

2.2 Data Processing

First, The study load the dataset and filter for wheat-related data. It is essential to extract records pertinent to wheat from the dataset. The dataset includes a diverse array of crop types. This experiment filters the data to retain only those rows where "Crop" is designated as "wheat"r to focus specifically on wheat." This step ensures that the subject of analysis pertains to a single crop, thereby eliminating any potential confusion arising from the growing conditions and yield data associated with different crops. The subsequent step involves the processing of rainfall and temperature data. These continuous numerical variables are converted into discrete intervals to enhance the analysis of the impacts of varying rainfall and temperature on wheat yield. This transformation facilitates a more effective comparison of yield performance across different climatic conditions. In the rainfall classification, the data collected in this experiment are categorized within 100 mm to 1000 mm, with an interval subsection of 18 mm. This indicates that rainfall is divided into several categorical intervals, and distinct rainfall conditions are classified as separate categories(arms). Various classifications can be established by categorizing the rainfall data into specific bins (e.g., 100-118mm, 118-136mm, etc.). These intervals effectively convert continuous rainfall data into discrete categorical data. In the case of temperature compartmentalization, temperature data are similarly categorized into intervals of 0.5° C, spanning from 15° C to 40° C.

Consequently, the temperature is divided into several distinct ranges, each represented as a single interval (e.g., 15.0-15.5° C, 15.5-16.0° C, etc.). Once the classification of rainfall and temperature has been completed, the corresponding wheat yield is calculated based on the rainfall or temperature within each specified interval. The critical step in this process involves categorizing the data into predefined intervals and tallying the output for each group. All yield data corresponding to that interval are aggregated for every rainfall interval, and the average yield is calculated accordingly. A similar methodology is applied to the temperature intervals to determine the average yield within each range. The study use this information to serve as a foundation for preparing the algorithm for subsequent evaluation.

2.3 Experimental Method

In this experiment, each algorithm undergoes an initial evaluation consisting of ten trials. The horizon value is set at 100,000 for this preliminary assessment. The results from these ten trials are then presented in a table using Matplotlib to determine the algorithms' feasibility within the context of this dataset. This step also allows for the evaluation of the algorithm's efficacy. If this algorithm demonstrates subpar performance in comparison to other algorithms, it will be excluded from the final comparative analysis. The initial level value of 100,000 was chosen to give each algorithm enough iterations to demonstrate its ability to learn and adapt. This allows an initial evaluation of algorithm performance without overstretching computing resources.

Subsequently, this experiment will perform a comparative analysis of the ultimately selected superior algorithm. In this final assessment, the horizon value is fixed at 1,000,000, and each algorithm undergoes testing for 100 iterations. It is reasonable to select a larger horizontal value for the final comparative analysis. This horizon ensures that each algorithm has ample opportunity to converge to the best solution, providing a more robust assessment of long-term performance and stability under different environmental conditions.Loop 100 times is to prevent some abnormal values due to accidental factors. I will average the 100 trials under the current horizon. This reduces errors and results in a more efficient assessment.The mean and standard deviation of these 100 trials in each round are calculated to generate the error bars for the algorithms. The error bars for each algorithm are then plotted on a graph to facilitate a clearer comparison of the distinctions among different algorithms.

2.4 Experimental Correlation Algorithm

The study utilizes four of the most widely recognized Multi-Armed Bandit (MAB) algorithms: the Explore Then Commit (ETC) algorithm, the Upper Confidence Bound (UCB) algorithm, the Asymptotically Optimal UCB algorithm, and the Thompson Sampling (TS) algorithm.

2.4.1 Explore Then Commit（ETC)

As shown in Table 1, the ETC algorithm is mainly divided into two stages: exploration and submission. During the exploration phase, the average reward for each arm is estimated by uniformly randomly sampling each type. The length of the exploration phase is where m represents the number of samples for each type and k indicates the number of types. For each type, m samples are randomly drawn from their corresponding reward list, and the average reward of these samples is calculated. In the submission stage, the type with the highest average reward (the optimal type) is selected, and only this type is chosen for the experiment in the submission stage. The length of the commit phase is, where T is the total experiment time. In this experiment, I set the exploration part mK to 0.1horizon. For most multi-armed bandits algorithm problems, the 10% exploration phase provides enough samples for each option to help the algorithm understand the performance of each option. This provides a relatively accurate feedback basis for the subsequent utilization stage, and improves the performance of the overall algorithm.

Table 1. Explore-Then-Commit Algorithm(ETC).

Explore Then Commit（ETC)

1.Input:m ∈ {1, . . . , ⌊T/K⌋}
2.for t = 1 → mK do
3. Select arm $a_t = (t \bmod K) + 1$ and observe reward R_t
4.end for
5.Calculate empirical mean reward for each arm a as:
$\hat{\mu}_a(mK) := \frac{\sum_{t=1}^{mK} R_t I\{a_t = a\}}{N_a(mK)}$
6.for t = mK + 1 →T do
7. Pull arm a: = arg $\max_{a\in[K]} \hat{\mu}_a(mK)$ (Commit)

2.4.2 Upper Confidence Bound（UCB）

As shown in Table 2, the Upper Confidence Bound (UCB) algorithm effectively integrates the dual objectives of development—selecting an arm with a high average reward—and exploration—attempting to select arms that are infrequently chosen. It calculates a confidence upper bound (UCB) for each arm, derived from two components: the average reward (utilization) associated with the arm and an additional term that encourages exploration for those arms that have not been frequently selected. The algorithm chooses the arm exhibiting the highest UCB value at each iteration. This approach guarantees the utilization of both arms that exhibit higher average rewards while simultaneously facilitating the exploration of less frequently selected rewards. This is particularly important as these lesser-chosen arms often yield valuable insights.The corresponding reward is observed by following the selection of an arm, and the count of selections for that specific arm is incremented. The process then continues to repeat. Over time, this algorithm progressively focuses on those options that provide higher rewards while periodically exploring choices to ensure no potentially superior alternatives are overlooked. This study set B to 2 here in order to adjust the size of this uncertainty term so that we can both maximize the known reward and explore potentially high reward options when choosing[15].

Table 2. Upper Confidence Bound(UCB).

Upper Confidence Bound（UCB）
1.Input：Horizon T, $\ell = 4$, δ
2.for each arm a ∈ [K], B = 2, $\delta = \frac{1}{T^2}$
3.for t = 1 → T do
4. Select arm $a_t = \text{argmax}_{a\in[k]} U_a(t-1, \delta)$
5. Observe reward R_t and update for a ∈ [K]:
$N_a(t) = N_a(t-1) + I\{a_t = a\}$ $\hat{\mu}_a(t) = \frac{N_a(t-1)\hat{\mu}_a(t-1)+R_t I a_t=a\}}{N_a(t)}$ $U_a(t, \delta) = \hat{\mu}_a(t) + \frac{B}{2}\sqrt{\frac{\ell \log(1/\delta)}{N_a(t)}}$
6.end for

2.4.3 Asymptotically Optimal UCB Algorithm

As shown in Table 3, based on the traditional Upper Confidence Bound (UCB) algorithm, the asymptotic UCB method incorporates a logarithmic term $\log(1+t\log(t)^2)$ to replace the original $\log(1/\delta)$. This term adds an exploration boundary that changes over time. By embedding the dependence of t into the logarithmic term, the algorithm can better handle the uncertainty that comes with growing over time. Increase of the term $\log(1+t\log(t)^2)$ is relatively slow over time, ensuring that the algorithm can smoothly reduce the exploration of options with greater uncertainty[16]. The remaining steps align with those of the conventional UCB algorithm. Again, set the value of B to 2.

Table3. Asymptotically Optimal UCB.

Asymptotically Optimal UCB
1.Input：Horizon T, $\ell = 4$
2.for each arm $a \in [K]$, $B = 2$, $\delta = \frac{1}{T^2}$
3.for $t = 1 \to T$ do
4. Select arm $a_t = \text{argmax}_{a\in[k]} U_a(t-1, \delta)$
5. Observe reward R_t and update for $a \in [K]$:
$N_a(t) = N_a(t-1) + I\{a_t = a\}$ $\quad \hat{\mu}_a(t) = \frac{N_a(t-1)\hat{\mu}_a(t-1)+R_t I a_t=a\}}{N_a(t)}$ $U_a(t, \delta) = \hat{\mu}_a(t) + \left(\frac{B}{2}\right)\sqrt{\frac{2\log(1+t\log(t)^2)}{N_a(t)}}$
6.end for

2.4.4 Thompson Sampling（TS）

As shown in Table 4, the TS algorithm effectively balances exploration and exploitation by leveraging the probability distribution of the arms during selection. The core principle involves randomly selecting arms based on current beliefs, represented by the posterior distribution, thereby facilitating natural exploration. Assuming that the returns from each arm follow a normal distribution, in each iteration, the parameters (mean and variance) for each arm are updated according to the observed returns, resulting in an updated posterior distribution. A value is sampled from this posterior distribution for each arm; subsequently, the arm with the highest sampled value is chosen. This methodology seamlessly integrates exploration and exploitation, as lower probabilities of exploration arise from more significant uncertainty. As more data becomes available, the posterior distributions are refined, reducing uncertainty and favouring those arms associated with higher returns—thereby enhancing utilization.

Table4. Thompson Sampling(TS).

Thompson Sampling（TS）
1.Input Horizon T

2.for each arm $a \in [K]$ initialize $\hat{\mu}_a = \mu_0, \hat{\sigma}_a^2 = \sigma_0^2$
3.for $t = 1 \to T$ do
4. for each arm $a \in [K]$:
Sample $\tilde{\mu}_a(t) \sim N(\hat{\mu}_a(t), \hat{\sigma}_a^2(t))$
5. select arm:$a_t = \arg\max_{a \in [K]} \tilde{\mu}_a(t)$
6. Observe reward R_t for arm a_t and update for a_t
7. $N_a(t) = N_a(t-1) + I\{a_t = a\}$ $\hat{\mu}_a(t) = \frac{N_a(t-1)\hat{\mu}_a(t-1) + R_t I a_t = a\}}{N_a(t)}$ $\hat{\sigma}_{a_t}^2(t) = \frac{1}{N_{a_t}(t)}$
8.end for

2.5 Algorithm Evaluation Method

This experiment incorporates the cumulative regret value concept to evaluate various algorithms' performance. This metric represents the total difference in returns between the arm selected by the algorithm and the optimal arm in each round.

$$R(t) = \sum_{s=1}^{T} (\mu^* - \mu_{a_s})$$

The lower the cumulative regret, the closer the algorithm's performance approaches the optimal choice. This indicates that the algorithm can converge to the optimal solution more rapidly. Therefore, an algorithm exhibiting lower cumulative regret over fewer rounds demonstrates greater efficiency in exploration.

In addition to calculating cumulative regret values. In the comparison experiment of algorithms, I also calculated the mean and variance of 100 trials of each algorithm in each round, so as to display the errorbar of each algorithm in the graph. It can represent the range of possible deviations above and below the estimate of a data point, helping us better understand the volatility and stability of the results.

3. Results and Analysis

3.1 Results of the Algorithm in the Context of Varied Rainfall Data for Wheat

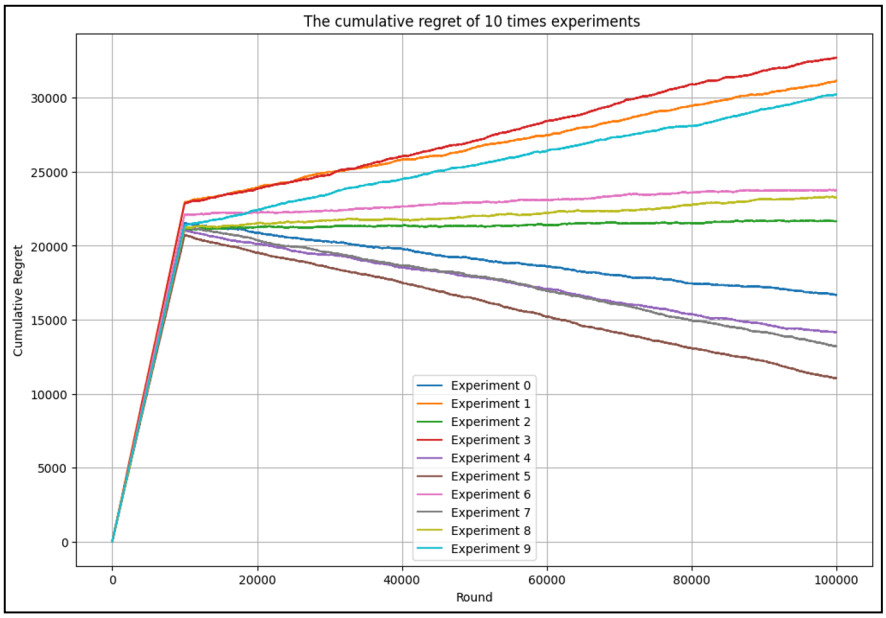

Figure1. Explore Then Commit(ETC).

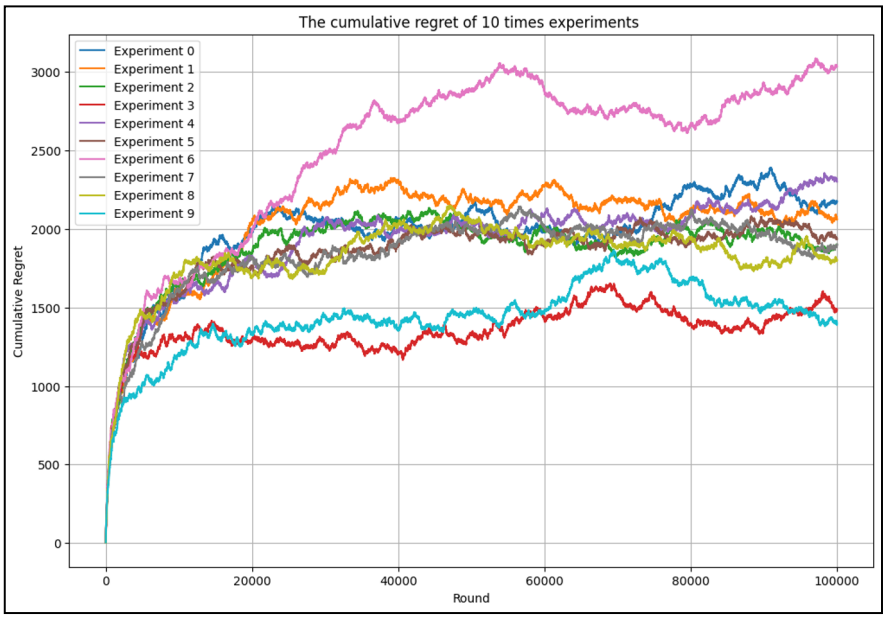

Figure2. Upper Confidence Bound(UCB).

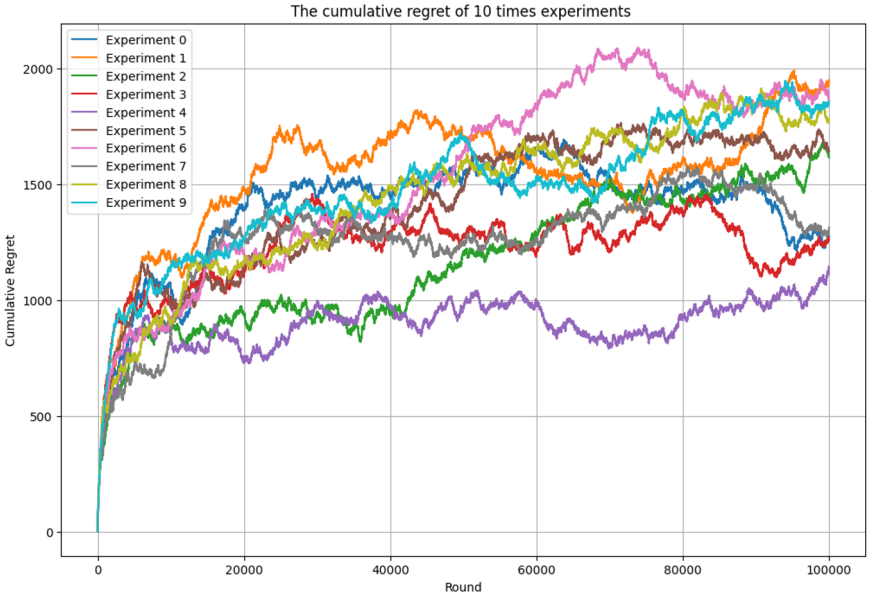

Figure3. Asymptotically Optimal UCB.

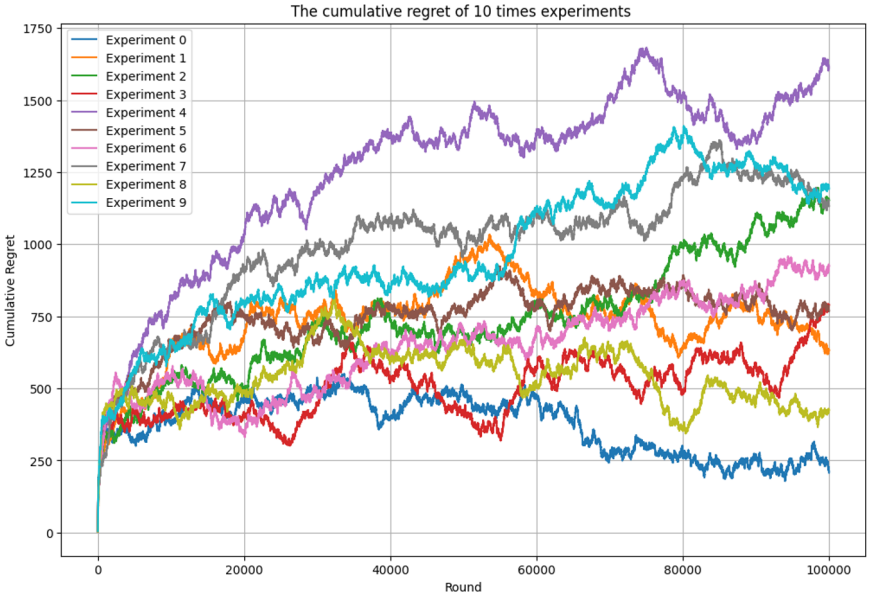

Figure4. Thompson Sampling(TS).

In the algorithm's preliminary exploration tests, the cumulative regret value associated with the ETC algorithm demonstrates a significant disparity across ten trials and is notably higher than that of other algorithms, as illustrated in Figures 1-4. This finding indicates that ETC exhibits considerable instability and demonstrates inferior predictive performance.Conversely, the total regret values for the other three algorithms eventually level off with only slight differences. This experiment provides a broader range of extra

tests on the UCB algorithm, Asymptotically Optimal UCB, and TS algorithm, considering that some errors could be due to different factors.

3.1.1 Algorithm Comparison Experiment

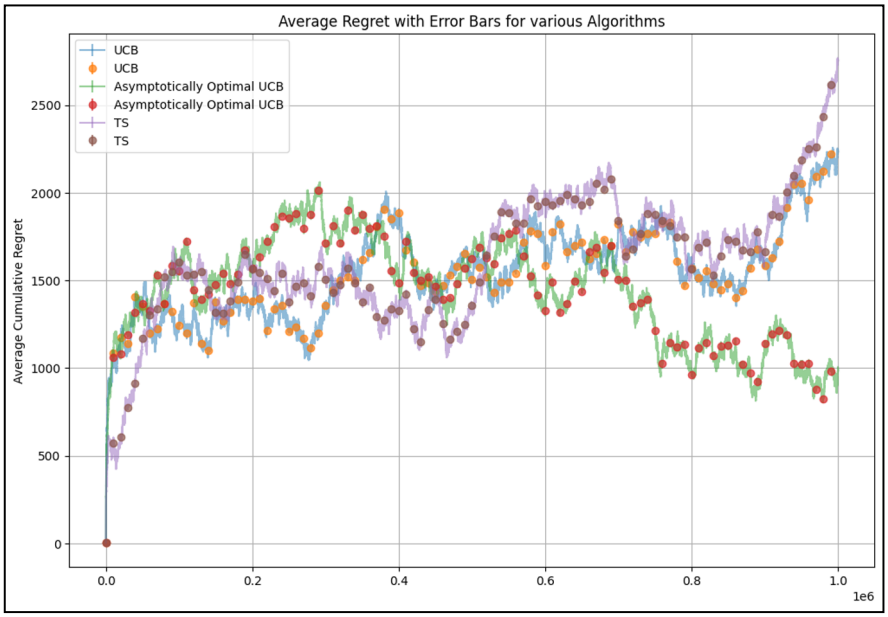

Figure5. Algorithm Comparison Experiment.

The study can observe that the Upper Confidence Bound (UCB) algorithm and the Asymptotically Optimal UCB algorithm in Figure 5 exhibit similar performance characteristics. Their cumulative regret values increase rapidly during the initial phase but gradually stabilize over time. The confidence boundary inherent in the UCB algorithm makes it more inclined towards "cautious" exploration. Although it may initially select a suboptimal strategy, the cumulative regret tends to stabilize over time, suggesting that it can effectively converge toward the optimal strategy. Asymptotically Optimal UCB excels in this regard, exhibiting markedly faster convergence and reduced cumulative regret over time. The performance of the TS algorithm is also similar to that of the UCB algorithm. Although the TS algorithm performs well in certain stages (particularly at the onset of the experiment), its cumulative regret value fluctuates significantly. This is because TS makes more arbitrary choices when uncertain, resulting in more dispersed selection strategies. The TS algorithm has the potential to gradually converge towards the optimal strategy. However, experimental results indicate a lack of favorable convergence trends over time. This can be attributed to a non-negligible probability of exploring non-optimal arms in later trials, suggesting that its stability in the wheat rainfall experiment is relatively inadequate.

3.2 Results of the Algorithm in the Context of Varied Temperature Data for Wheat

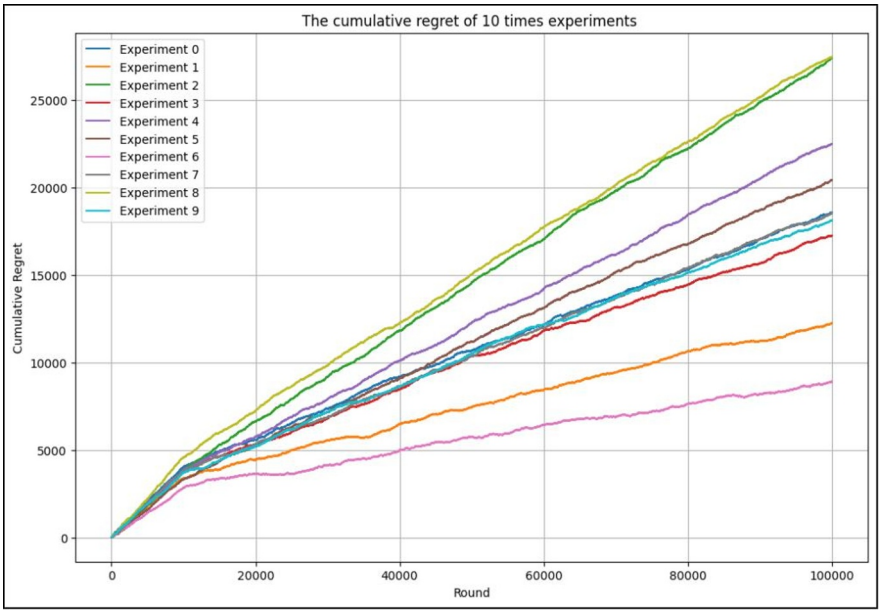

Figure6. Explore Then Commit(ETC).

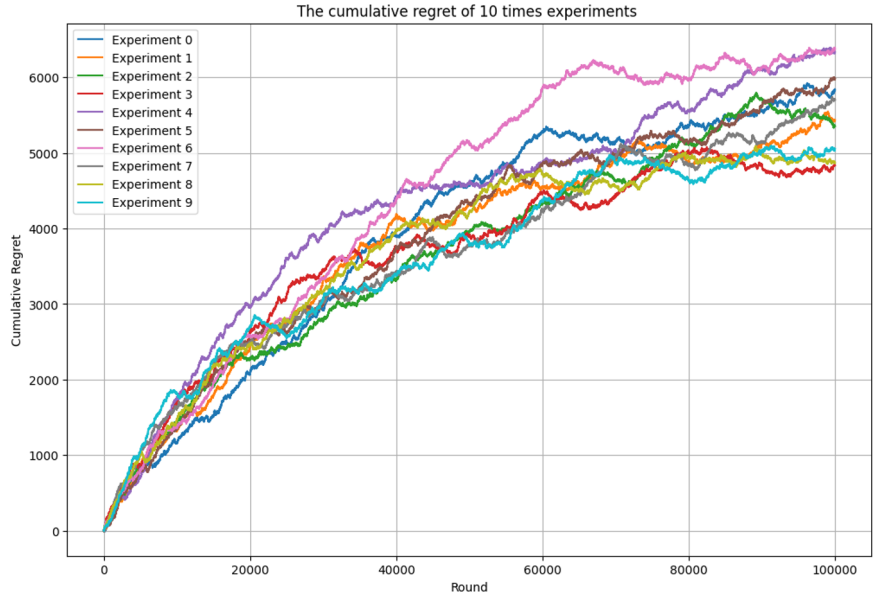

Figure7. Upper Confidence Bound(UCB).

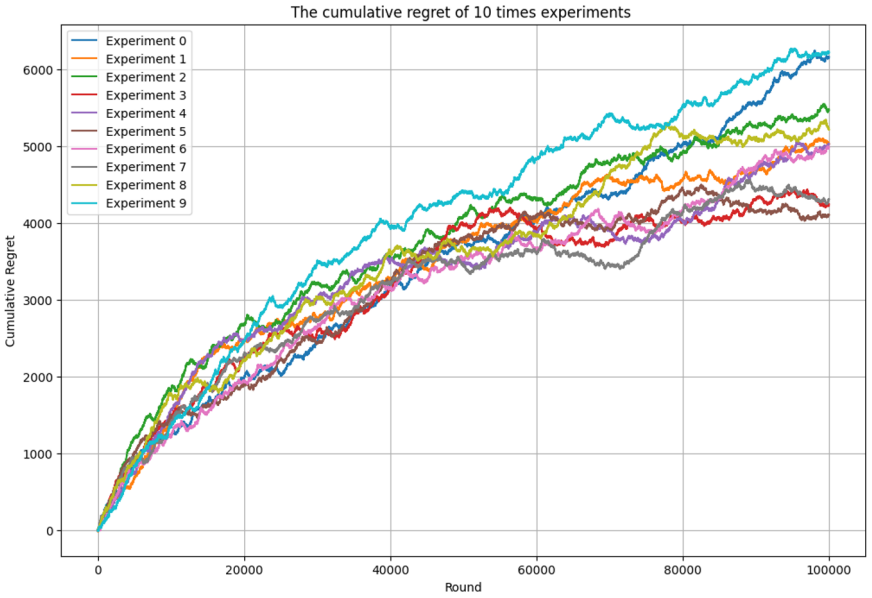

Figure8. Asymptotically Optimal UCB.

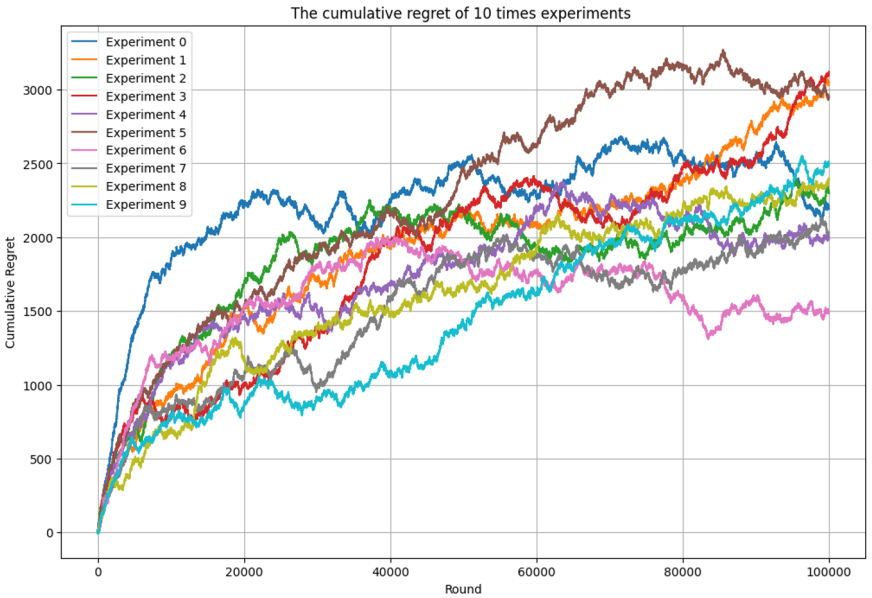

Figure9. Thompson Sampling(TS).

Figures 6-9 indicate that the cumulative regret value of the ETC algorithm for temperature data rises significantly as the horizon increases in ten tests and is considerably higher than that of other algorithms. It has been proved that in temperature data, it is difficult for the ETC algorithm to find the optimal arm, which reflects the poor prediction effect. The cumulative regret value trend of the other three algorithms in ten experiments remains consistent with that of the rainfall data, and the disparity in value is not particularly

prominent. Similarly, The study will subject these three algorithms to a more excellent range of values for more tests to explore the optimal algorithm and its distinctions.

3.2.1 Algorithm Comparison Experiment

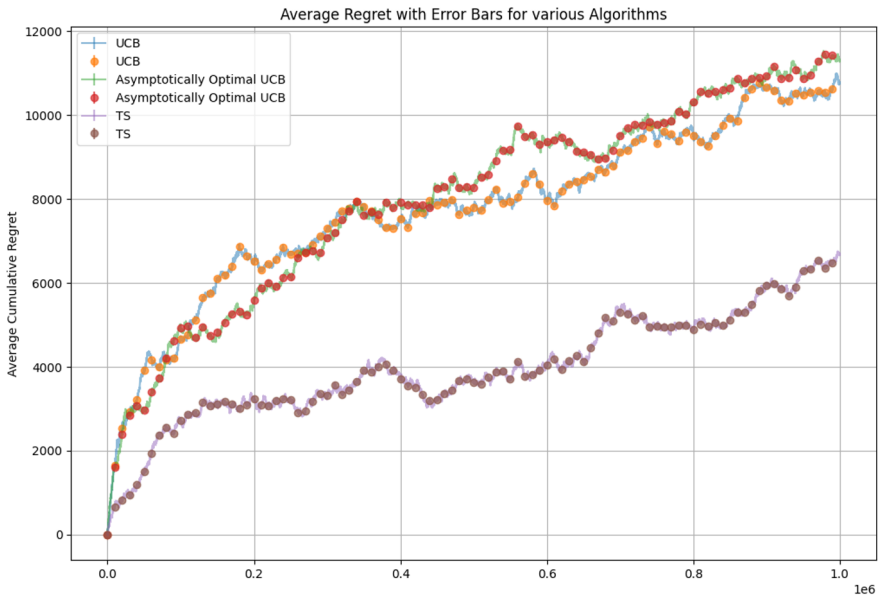

Figure10. Algorithm Comparison Experiment.

The study observe from Figure 10 that the UCB algorithm and the Asymptotically Optimal UCB algorithm align in most instances. The cumulative regret value rose rapidly in the initial stage and gradually increased as the number of experiments increased. During the initial phase of the experiment, the TS algorithm's performance was markedly superior to that of the other two algorithms. This is because its exploration mechanism is more flexible, and it can rapidly discover through Bayesian updating which temperature (arm) is likely to yield a higher reward. In contrast to precipitation, temperature exhibits lower volatility in a specific region. This allows TS to promptly focus on the relatively stable and high-yield range in the early exploration stage without spending considerable time on ineffective exploration. This feature empowers the TS algorithm to construct a superior temperature-yield relationship model quickly, thereby minimizing the cumulative regret value. As the number of tests increases, the cumulative regret value of TS does not flatten out. This is because the algorithm exhibits certain randomness when confronted with long-term decisions, resulting in its inability to balance exploration and utilization more precisely than UCB. Overall, however, the effect of the TS algorithm on wheat temperature data is significantly better than that of the other two algorithms.

4.Conclusion

In this study, The study investigate the application of multi-armed bandit (MAB) algorithms—specifically, the Explore-Then-Commit (ETC), Upper Confidence Bound (UCB), Asymptotically Optimal UCB, and Thompson Sampling (TS)—to optimize

agricultural decision-making related to wheat production under varying temperature and precipitation conditions. Our experiments clarify both the advantages and constraints of these algorithms in tackling environmental uncertainty, as well as their effects on cumulative regret.

The ETC algorithm shows poor performance in handling both temperature and rainfall. One of its significant drawbacks is that its exploration phase is fixed and ends once the exploration period is over. In experiments related to temperature or precipitation, the complexity of environmental changes might require continuous update strategies. The ETC algorithm is unable to adapt to new data after the exploration period, causing it to miss subsequent optimization opportunities.

In experiments examining the total precipitation during the growth period of wheat, both UCB and Asymptotically Optimal UCB exhibit low cumulative regrets and consistently demonstrate exceptional performance in managing the uncertainties associated with rainfall. Although the TS algorithm initially demonstrates robust performance, it faces challenges as the experiment progresses, resulting in a significant increase in its cumulative regrets during the later stages. This underscores the algorithm's difficulty in maintaining consistent decision-making within highly variable precipitation environments.

When investigating the average temperature during wheat growth, the Thompson Sampling (TS) algorithm demonstrates significantly superior outcomes to the Upper Confidence Bound (UCB) and Asymptotically Optimal UCB. The TS algorithm can rapidly identify the potentially optimal arm (temperature interval) by sampling from the probability distribution, thereby effectively reducing cumulative regret within a short timeframe. However, its performance remains considerably lower than that of the other two algorithms in subsequent processes.

Our research highlights the significant potential of intelligent algorithms in enhancing agricultural decision-making. By employing the Multi-Armed Bandit (MAB) algorithm, farmers can optimize their wheat cultivation strategies across various environmental conditions, particularly in response to uncertainties stemming from climate change. UCB and TS algorithms are particularly auspicious in precision agriculture, offering a potent framework for real-time decision-making.

Future studies could contextualize diverse environmental variables to construct a tailored TS algorithm or perpetually enhance the UCB algorithm for better optimization and prediction of the most appropriate environmental conditions for agricultural products. This would facilitate agricultural practitioners' more effective adjustment of their existing strategies and higher yields.

References

1. Yang, Y., Tilman, D., Jin, Z., Smith, P., Barrett, C. B., Zhu, Y. G., ... & Zhuang, M. (2024). Climate change exacerbates the environmental impacts of agriculture. Science, 385(6713), eadn3747.
2. Kurniawan, D. A., & Santoso, A. Z. (2020). Pengelolaan sampah di daerah sepatan kabupaten tangerang. ADI Pengabdian Kepada Masyarakat, 1(1), 31-36.
3. Delgado, J. A., Short Jr, N. M., Roberts, D. P., & Vandenberg, B. (2019). Big data analysis for sustainable agriculture on a geospatial cloud framework. Frontiers in Sustainable Food Systems, 3, 54.
4. Burkett, V. R., Suarez, A. G., Bindi, M., Conde, C., Mukerji, R., Prather, M. J., ... & Nyambod, E. (2015). Point of departure. In Climate Change 2014 Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects (pp. 169-194). Cambridge University Press.

5. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.
6. Sottocornola, G., Nocker, M., Stella, F., & Zanker, M. (2020, March). Contextual multi-armed bandit strategies for diagnosing post-harvest diseases of apple. In Proceedings of the 25th international conference on intelligent user interfaces (pp. 83-87).
7. Baudry, D., & Gautron, R. Risk-Aware Bandits for Best Crop Management. In ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists.
8. Tekin, C., & Liu, M. (2012). Online learning of rested and restless bandits. IEEE Transactions on Information Theory, 58(8), 5588-5611.
9. Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.
10. Gautron, R., Maillard, O. A., Preux, P., Corbeels, M., & Sabbadin, R. (2022). Reinforcement learning for crop management support: Review, prospects and challenges. Computers and Electronics in Agriculture, 200, 107182.
11. Uyeh, D. D., Hiablie, S., Park, T., Bassey, B. I., Mallipeddi, R., Woo, S., ... & Ha, Y. (2022). Optimal sensors placement in controlled environment agriculture using a reinforcement learning approach. In 2022 ASABE Annual International Meeting (p. 1). American Society of Agricultural and Biological Engineers.
12. Stetter, C., Huber, R., & Finger, R. (2024). Agricultural land use modeling and climate change adaptation: A reinforcement learning approach. Applied Economic Perspectives and Policy.
13. Auer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem.
14. Agrawal, S., & Goyal, N. (2012, June). Analysis of thompson sampling for the multi-armed bandit problem. In Conference on learning theory (pp. 39-1). JMLR Workshop and Conference Proceedings.
15. Auer, P. (2002). Using upper confidence bounds for exploration in reinforcement learning. Proceedings of the 19th International Conference on Machine Learning (ICML), 21-28.
16. Garivier, A., & Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. Proceedings of the 24th Annual Conference on Learning Theory (COLT), 359-376.