Estimating the total number of active wheat harvesters using big data of GNSS trajectories in China

Jiawei Xu[1,2], Yihui Li[3], Yingkuan Wang[4,5*], Caicong Wu[1,2*]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
2. *Key Laboratory of Agricultural Machinery Monitoring and Big Data Applications, Ministry of Agriculture and Rural Affairs, Beijing 100083, China*;
3. *College of Science, China Agricultural University, Beijing 100083, China*;
4. *Academy of Agricultural Planning and Engineering, Ministry of Agriculture and Rural Affairs, Beijing 100125, China*;
5. *Chinese Society of Agricultural Engineering, Beijing 100125, China*)

Abstract: China plants approximately 20.3 million hm^2 of winter wheat annually. During the recent one-month harvesting period, hundreds of thousands of combine harvesters participated in wheat harvesting from south to north. However, the total number of active harvesters remains a challenge, restricting government policy-making and industry analysis. This study proposed a nonparametric bootstrap estimation model based on big data to dynamically infer the total number of active agricultural machines by analyzing the spatio-temporal trajectories of harvesters. Through Monte Carlo simulation experiments, the performance of four nonparametric bootstrap methods was systematically evaluated from dimensions such as bias, mean squared error, and coverage probability. The results show that the bias-corrected and accelerated bootstrap method (BCa) performs best and was selected as the 95% confidence interval estimation method. The 95% confidence intervals for the total number of active harvesters in 2021, 2022, and 2023 are [447 223, 456 387], [441 708, 447 625], and [436 873, 440 608], respectively, providing a quantitative basis for regulatory supervision and capacity planning in the agricultural machinery industry.

Keywords: wheat harvester, big data, total active number, confidence interval
DOI: 10.25165/j.ijabe.20251804.9151

Citation: Xu J W, Li Y H, Wang Y K, Wu C C. Estimating the total number of active wheat harvesters using big data of GNSS trajectories in China. Int J Agric & Biol Eng, 2025; 18(4): 195–199.

1 Introduction

China grows about 20.3 million hm^2 of winter wheat every year, with the average planting area per household being only about 0.3 hm^2, and the north-south planting span is about 950 km. Therefore, China adopts large-scale cross-regional harvesting operations. During the harvesting period of nearly one month, harvesters move from south to north, and complete all winter wheat harvesting day by day. The Ministry of Agriculture and Rural Affairs of China announced that the total number of harvesters in stock is about 1.6 million units, and the active number is about 650 000 units, which are based on the manufacturer's sales volume, the number of machinery purchase subsidies, and the empirical estimates. However, the actual total number of active harvesters should be lower than this value since sales volume is not equal to the active number. For example, many harvesters broke down or were even scrapped, but the government failed to keep statistics on time. While the sowing area is determined, if there are too many active harvesters, the harvesting income will be too low. On the contrary, if the total number of active harvesters is too small, it will affect the timely harvesting of wheat, eventually causing food losses. Therefore, maintaining a reasonable number of active harvesters is not only related to the operating income of each operator, but also to the timely harvesting of wheat. Beyond ensuring an optimal number of machines, harvest outcomes are also constrained by other factors. For instance, current scheduling often relies on experience, leading to inefficient resource allocation[1]. Furthermore, meteorological hazards such as low-temperature stress significantly impact winter wheat growth and yield[2]. These hazards can cause the final harvested area to be smaller than the planted area by as much as 13%, which leads to overestimations of total production if not properly accounted for[3].

There are reasons why this problem (estimation of the total number of active harvesters) has not been solved. In the early stage, it was difficult to collect a sufficient sample size for statistical analysis. For example, a 5% sample size was about 25 000 units of harvesters, which was too large for a researcher to track their harvesting process and area without GNSS-enabled terminals. Therefore, it was difficult to establish a corresponding statistical model to accurately measure the total active number. Now, China's Agricultural Machinery Operation Big Data System[4] based on GNSS-enabled terminals and has accessed to 280 000 units of harvesters (hereinafter "GNSS harvester"). Based on the system, the operating area can be accurately calculated for each GNSS harvester, laying the data foundation for the estimation of the total number of active wheat harvesters.

In the field of estimating the total number of active vehicles, there is no relevant literature in the agricultural sector. However, there have been several cases in the field of urban traffic. Torok et al.[5] modeled the relationship between per capita GDP and vehicle ownership growth based on Hungarian passenger car data using the Gompertz function. Javid et al.[6] analyzed the development trend of

Received date: 2024-06-14 **Accepted date:** 2025-07-09
Biographies: Jiawei Xu, PhD, research interest: agricultural machinery big data analysis, Email: xujiaweiciee@cau.edu.cn; **Yihui Li**, Bachelor, research interest: agricultural machinery big data analysis, Email: 20203170102221@cau.edu.cn.
***Corresponding author: Yingkuan Wang**, PhD, Researcher, research interest: agricultural engineering, intelligent agricultural equipment. Academy of Agricultural Planning and Engineering, Ministry of Agriculture and Rural Affairs, Beijing 100125, China. Tel: +86-10-59197088, Email: wangyingkuan@163.com; **Caicong Wu**, PhD, Professor, research interest: the navigation and big data mining of agricultural machinery. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Tel: +86-13810521813, Email: wucc@cau.edu.cn.

electric vehicle ownership in California through multivariate logistic regression. Zhang et al.[7] utilized Beijing's 2017 travel data, heat maps, and POI data to analyze the impact of accessibility on household car ownership through the Gradient Boosting Decision Trees (GBDT) algorithm. However, these studies have two main shortcomings: firstly, the prediction scope is usually limited to a specific city or region; secondly, the amount of data relied upon is still relatively small, and the advantages of big data have not been fully utilized.

In the field of statistics, the Law of Large Numbers (LLN) ensures that with a large number of repeated experiments or a sufficiently large sample size, the sample mean will approach the population mean, providing a theoretical foundation for the accuracy of estimations and the reliability of predictions[8]. It has been widely applied in fields such as computer technology[9], social sciences[10], and health insurance[11]. Interval estimation not only provides parameter estimates but also quantifies the uncertainty of those estimates by giving a confidence interval, making the estimation more comprehensive and reliable. In agriculture, parameter estimation has been applied to rice yield estimation[12,13], crop straw resource estimation[14], and the total area of salt-affected soils[15], and all these studies have achieved good results.

2 Materials and methods

2.1 Technical route

This study first conducted data cleaning on trajectory data, followed by field-to-road segmentation and area calculation, to obtain information on the daily operating area of each combine harvester and its corresponding province, city, and county within the platform from 2021 to 2023. Based on the Law of Large Numbers and four non-parametric Bootstrap models, the number of active wheat combine harvesters was estimated. The specific technical route is shown in Figure 1.

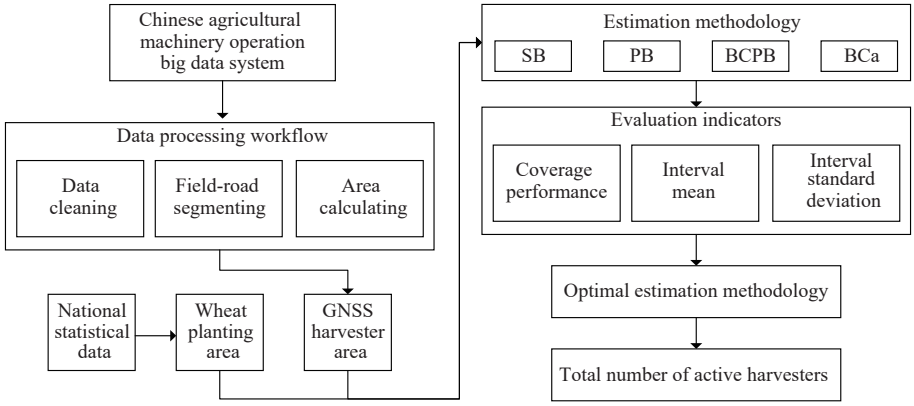


Figure 1 Technical route for calculating the total number of active harvesters

2.2 Dataset and data processing

2.2.1 Original dataset

The dataset of this study comes from the Chinese Agricultural Machinery Operation Big Data System from 2021 to 2023. This system stores trajectory data of the harvesters working in nine major wheat-producing provinces of China (Hebei, Henan, Shanxi, Shandong, Anhui, Hubei, Shaanxi, Jiangsu, and Sichuan). The harvesters' trajectory data spans from May 1 to June 25 of each year. The fields of the trajectory data include parameters such as the harvesters' ID, operation date, operation time, longitude, latitude, speed, and direction. 95% of the trajectories have a reporting frequency of no more than 5 s. Among them, the wheat planting area was obtained from the official data published by the provincial governments of the major producing areas.

2.2.2 Data preprocessing

1) Data cleaning. This study performed preprocessing on the trajectory data, such as noise smoothing, drift point removal, and irregular velocity point elimination, in order to remove the abnormal data.

2) Field-road segmenting. The density-based spatial clustering algorithm (DBSCAN) was used to realize the field-road classification of the trajectory, as shown in Figure 2, and its accuracy can reach 96.01%, with an F1 score of 95.60%[16].

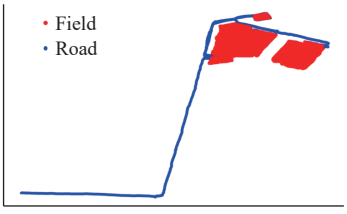


Figure 2 Field-road segmentation based on DBSCAN

3) Area calculating. The area calculation method used in this study is the grid key point method, with the bottom grid size set to 1 m×1 m[17].

2.2.3 Final dataset

Finally, the detailed operation data of each harvester on each farmland are able to be obtained, including the farmland ID, operation area, operation start and end time, latitude and longitude of the center point, and the province, city, and county where the harvester worked.

In order to eliminate the abnormal operating area, this study set a threshold for the harvesting area according to the investigation. Minimum total harvesting area: During one season, if the total harvesting area is less than 2.02 hm^2, the harvester would be excluded. Maximum harvesting area per farmland: If the harvesting area of a single farmland is larger than 47.23 hm^2, the area data would be excluded. All excluded anomalous data underwent manual verification. Due to poor data quality, they could not be utilized.

According to the above thresholds, the number of harvesters excluded in 2021, 2022, and 2023 is 4830, 3144, and 5443, respectively.

2.3 Estimating total number of active harvesters and its confidence interval

2.3.1 Confidence interval estimate

Bootstrap is a scientific statistical method that estimates the sampling distribution of a statistic by repeatedly sampling the original sample data. As shown in Figure 3, assume that $\{x_1, x_2, x_3, \ldots, x_n\}$ is a random sample of size n (population sample) drawn from a process, where n is an integer greater than or equal to 1000, and $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$ represents a sample of size n drawn from the

original sample with replacement. Therefore, there are n^n possible resampling methods.

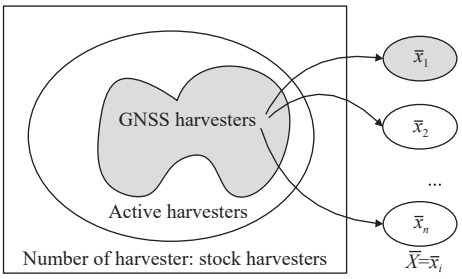


Figure 3 Schematic diagram of Bootstrap sampling

According to different sampling methods, Bootstrap methods can be divided into two types: parametric Bootstrap and non-parametric Bootstrap. Compared with the parametric Bootstrap method, the core advantage of the non-parametric Bootstrap method lies in its non-parametric nature, that is, it does not need to make any assumptions about the underlying distribution of the data. At the same time, it can make full use of the information in the sample data, and the estimation results are more accurate. This feature makes the non-parametric Bootstrap method particularly suitable for actual data analysis scenarios where it is difficult to meet the distribution assumptions required by traditional parametric statistical methods. In our task, the number of active harvesters is estimated by sampling using non-parametric Standard Bootstrap[18,19] and its three modified Bootstrap methods, including Percentile Bootstrap[20], Biased-Corrected Percentile Bootstrap[21], and Bias-Corrected and Accelerated Bootstrap[22]. Then, the accuracy of the four methods was evaluated, and the result of the optimal method was selected as the estimation result of the number of active harvesters.

2.3.2 Evaluation index

When choosing the optimal method among the four estimation methods for estimating the total number of active harvesters from 2021 to 2023, it was crucial to consider the two key indicators of accuracy and precision. Accuracy and precision measure the ability of an estimation method to measure the true value of a feature and the variability of the measurements, respectively. In order to comprehensively evaluate the effects of the four methods, this study used Monte Carlo simulation research to numerically analyze the statistical properties of the four confidence intervals from three perspectives: coverage performance index, interval mean index, and interval standard deviation index.

3 Results and discussion

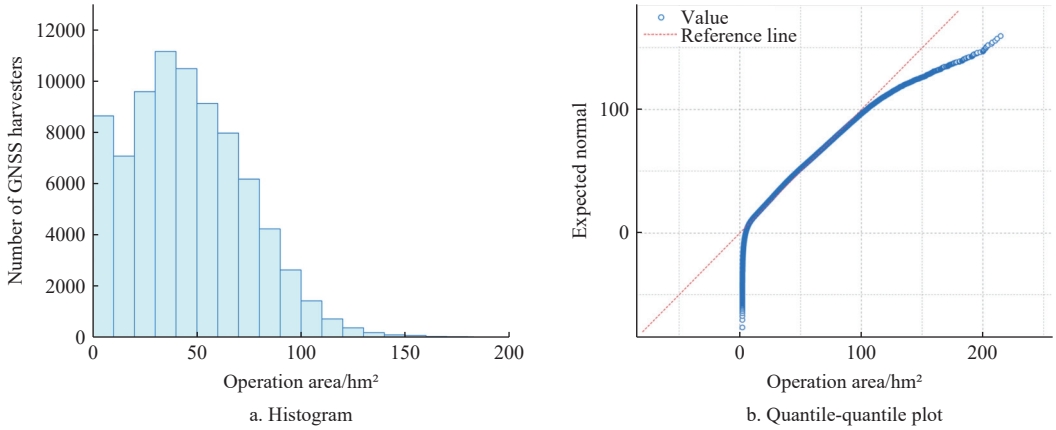
3.1 Average harvesting area of GNSS harvesters

Through data pre-processing, area calculation, and statistical analysis, the operation data of GNSS harvesters from 2021 to 2023 are listed in Table 1. It can be seen that as the number of GNSS harvesters increases year by year, the total operating area of GNSS harvesters has grown rapidly, but the average operating area has not increased much. In 2021, 2022, and 2023, the operating area of GNSS harvesters accounted for 5.17%, 12.89%, and 18.28%, respectively. These data show that the dataset established by the author has a large scale.

Table 1 2021-2023 GNSS harvester operating area statistical data

Year	Number of GNSS harvesters	Total area of wheat in the main producing provinces/hm²	Total operation area by GNSS harvester/ hm²	Average area of GNSS harvester/ hm²	Proportion of GNSS harvester operations
2021	23 356	20 117 410	1 039 853	44.52	5.17%
2022	57 314	20 262 110	2 611 074	45.56	12.89%
2023	80 188	20 57 510	3 702 141	45.50	18.28%

Figure 4 shows the histogram and QQ plot of the operating area of GNSS harvesters in 2023. It can be observed that the data distribution of the operating area of GNSS harvesters in 2023 significantly deviates from the characteristics of the normal distribution. To statistically verify this observation, this study further used the Kolmogorov-Smirnov test to evaluate the normality of the data from different dimensions. The results show that the operating area data of GNSS harvesters in 2023 do not follow a normal distribution, and the data in 2021 and 2022 also do not follow a normal distribution.



a. Histogram b. Quantile-quantile plot

Figure 4 Statistics of GNSS harvesters' operating area in 2023

After obtaining the GNSS harvesters' operating area data in China's nine main wheat-producing provinces from 2021 to 2023, this study screened the operating area of the newly added GNSS harvesters every year and calculated their average operating area (Table 2).

Through horizontal comparison, it was found that the average operating area of newly added GNSS harvesters in 2022 and 2023 showed a year-by-year downward trend within three years after purchase. Further vertical comparison revealed that the average operating area of newly purchased GNSS harvesters each year showed a downward trend compared with that of old harvesters. This may be because in China's main wheat-producing areas, the newly purchased GNSS harvesters did not show significant differences in work efficiency compared with the early batches of

Table 2 Average operating area of the newly added GNSS harvesters (hm²)

Year	2021	2022	2023
2021	57.25	52.74	46.90
2022	-	43.84	51.79
2023	-	-	38.76

old harvesters not equipped with GNSS terminals. The reason behind this phenomenon may be that the total number of harvesters in China has become saturated, making it difficult to fully leverage the high-performance advantages of new harvesters over old ones.

3.2 Estimation of harvester interval

Given the non-normal distribution of the data, traditional parametric statistical methods that rely on the assumption of normal distribution may not be suitable for this study. Therefore, this study turned to non-parametric statistical methods and selected the non-parametric Bootstrap method for further analysis. The non-parametric Bootstrap method does not rely on the specific distribution assumptions of the data, so it shows higher flexibility and applicability when dealing with non-normally distributed data.

By using the non-parametric standard Bootstrap method and three improvements, this study successfully estimated the number of active harvesters in the main wheat-producing areas from 2021 to 2023 and their 95% estimation interval. Combining the results obtained from these four methods shows that the total number of active harvesters from 2021 to 2023 and their estimated interval have a high degree of confidence (Table 3).

Table 3 Comparison of estimation methods for the active number of harvesters in 2023

Year	Number of GNSS harvesters	Method	95% confidence interval	
			Lower limit	Upper limit
2021	23 356	SB	447 271	456 535
		PB	447 106	456 846
		BCPB	447 245	456 749
		BCa	447 223	456 387
2022	57 314	SB	441 767	447 794
		PB	441 703	447 754
		BCPB	441 940	447 636
		BCa	441 708	447 625
2023	80 188	SB	436 939	440 507
		PB	436 753	440 694
		BCPB	436 907	440 588
		BCa	436 873	440 608

The coverage performance index, interval mean index, and interval standard deviation index obtained by the four estimation methods are shown in Figure 5.

After a comprehensive comparative analysis of the four methods, it was found that the Standard Bootstrap (SB) showed the lowest interval standard deviation but longer interval mean index and lower coverage performance index, potentially reducing result accuracy. The other methods had similar interval standard deviation and interval mean index, but the Bias-Corrected and Accelerated (BCa) Bootstrap method had a higher coverage performance index, indicating greater reliability.

Therefore, it can be concluded that the Bias-Corrected and Accelerated (BCa) Bootstrap method performs best in terms of accuracy and precision. Therefore, the BCa method was selected as the main method for estimating the total number of active harvesters from 2021 to 2023, and the final results are shown in Figure 6. The data in the table is the estimated value of the total number of active harvesters in the main wheat-producing areas from 2021 to 2023, calculated using the BCa method and the corresponding 95% estimate interval.

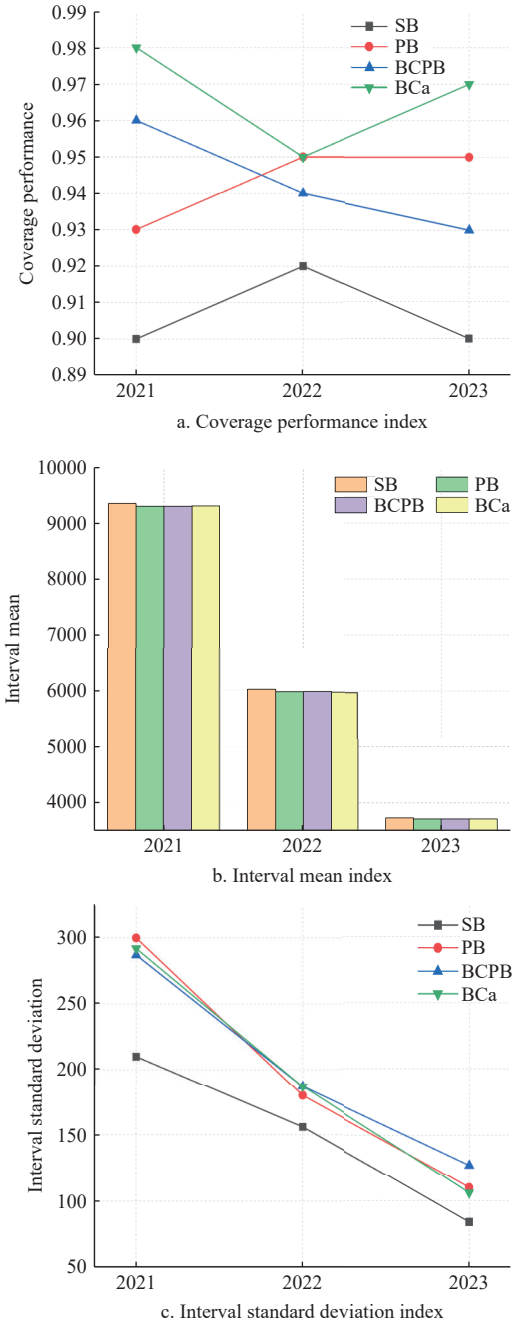


Figure 5 Performance comparison of four methods

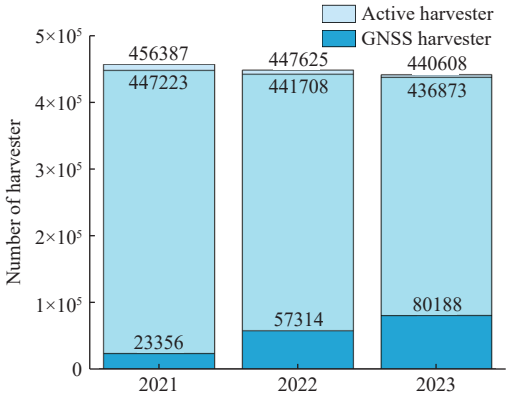


Figure 6 Estimated total number of active harvesters in 2021-2023

4 Conclusions

This study proposes a method for estimating the total number of active harvesters based on big data. Using GNSS harvester trajectory data as the foundation, this method accurately calculates the operational area of each GNSS harvester and then derives the average harvesting area of all GNSS harvesters. By integrating the total sown area of winter wheat in China's main wheat-producing provinces from 2021 to 2023, and utilizing the 95% Bootstrap confidence interval, this method provides confidence intervals for the number of active harvesters. The specific conclusions are as follows:

1) The operational area data of GNSS harvesters in China's main wheat-producing regions from 2021 to 2023 do not follow a normal distribution.

2) The 95% confidence interval values estimated by the four methods (SB, PB, BCPB, and BCa) are similar, demonstrating the feasibility of using these methods for estimating the number of active harvesters.

3) Newly purchased GNSS harvesters have not shown significant differences in work efficiency compared with older harvesters not equipped with GNSS terminals.

4) The total number of harvesters in China may have become saturated, making it difficult for new harvesters to fully leverage their high-performance advantages over old models.

5) From the three perspectives of coverage performance index, interval mean index, and interval standard deviation index, the four non-parametric estimation methods were evaluated, and it was found that BCa had the best effect.

6) The 95% confidence intervals for the total number of active harvesters from 2021 to 2023 estimated by the BCa method are [447 223, 456 387], [441 708, 447 625], and [436 873, 440 608].

Acknowledgements

This work was financially supported by the National Precision Agriculture Application Project (Grant/Contract No. JZNYYY001).

[References]

[1] Cao G Q, Ma B, Chen C, Ren B X, Hu C Z. Agricultural machinery cross-region scheduling optimization based genetic algorithm variable neighborhood search. Transactions of the CSAM, 2023; 54(10): 114–123.

[2] Chen J M, Zhang P Y, Liu J M, et al. Study on the impact of low-temperature stresson winter wheat based on multi-model coupling. Food and Energy Security, 2024; 13: e543.

[3] Hu J K, Zhang B, Peng D L, Huang J X, Zhang W J, Zhao B, et al. Mapping 10-m harvested area in themajor winter wheat-producing regions of China from 2018 to 2022. Scientific Data, 2024; 11: 1038.

[4] Wu C C, Li D, Zhang X Q, Pan J W, Quan L, Yang L L, et al. China's agricultural machinery operation big data system. Computers and Electronics in Agriculture, 2023; 205: 107594.

[5] Torok A. Prediction of vehicle ownership growth using Gompertz model, case study of Hungary. System Safety: Human-Technical Facility-Environment, 2022; 4(1): 164–169.

[6] Javid R J, Nejat A. A comprehensive model of regional electric vehicle adoption and penetration. Transport Policy, 2017; 54: 30–42.

[7] Zhang W J, Zhao Y J, Cao X J, Lu D M, Chai Y W. Nonlinear effect of accessibility on car ownership in Beijing: Pedestrian-scale neighborhood planning. Transportation Research Part D: Transport and Environment, 2020; 86: 102445.

[8] Karatzas I, Schachermayer W. A strong law of large numbers for positive random variables. Illinois Journal of Mathematics, 2023; 67: 517–528.

[9] Sirignano J, Spiliopoulos K. Mean field analysis of neural networks: A law of large numbers. SIAM Journal on Applied Mathematics, 2020; 80: 725–752.

[10] Schläpfer M, Dong L, O Keeffe K, Santi P, Szell M, Salat H, et al. The universal visitation law of human mobility. Nature, 2021; 593: 522–527.

[11] Buchmueller T C, Levy H G. The ACA's impact on racial and ethnic disparities in health insurance coverage and access to care: an examination of how the insurance coverage expansions of the Affordable Care Act have affected disparities related to race and ethnicity. Health Affair, 2020; 39: 395–402.

[12] Nodin M N, Mustafa Z, Hussain S I. Assessing rice production efficiency for food security policy planning in Malaysia: A non-parametric bootstrap data envelopment analysis approach. Food Policy, 2022; 107: 102208.

[13] Tiwari V, Thorp K, Tulbure M G, Gray J, Kamruzzaman M, Krupnik T J, et al. Advancing food security: Rice yield estimation framework using time-series satellite data & machine learning. PloS One, 2024; 19: e0309982.

[14] Aquino D, Del Barrio A, Trach N X, Hai N T, Khang D N, Toan N T, et al. Rice straw-based fodder for ruminants. Sustainable Rice Straw Management, 2020; pp.111-129. doi: 10.1007/978-3-030-32373-8_7.

[15] Negacz K, Malek Z, de Vos A, Vellinga P. Saline soils worldwide: Identifying the most promising areas for saline agriculture. Journal of Arid Environments, 2022; 203: 104775.

[16] Chen Y, Quan L, Zhang X Q, Zhou K, Wu C C. Field-road classification for GNSS recordings of agricultural machinery using pixel-level visual features. Computers and Electronics in Agriculture, 2023; 210: 107937.

[17] Xu J W, Kuang K M, Fu C, Wu C C. Calculation of the harvested acreage for wheat harvesters based on spatiotemporal trajectories and grid key points. Transactions of the CSAE, 2025; 41(11): 35–40.

[18] Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika, 1981; 68: 589–599.

[19] Efron B. The jackknife, the bootstrap and other resampling plans. SIAM, 1982; pp.13-19. doi: 10.1137/1.9781611970319.

[20] Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 1983; 37(1): 36–48.

[21] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1986; 1(1): 54–75.

[22] Efron B. Better bootstrap confidence intervals. Journal of the American Statistical Association, 1987; 82(397): 171–185.