IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 December 2023, accepted 24 January 2024, date of publication 26 January 2024, date of current version 5 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359055

METHODS

Development and Assessment of SPM: A Sigmoid-Based Model for Probability Estimation in Non-Repetitive Unit Selection With Replacement

SAMARTH GODARA[1], G. AVINASH[1], RAJENDER PARSAD[1], SUDEEP MARWAHA[1], MUKHTAR AHMAD FAIZ[2,3], AND RAM SWAROOP BANA[3]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India
[2]Department of Agronomy, Afghanistan National Agricultural Sciences and Technology University, Kandahar 3801, Afghanistan
[3]ICAR-Indian Agricultural Research Institute, New Delhi 110012, India

Corresponding author: Mukhtar Ahmad Faiz (maf.maher.anastu@gmail.com)

ABSTRACT Probability estimation plays a pivotal role across diverse domains, particularly in scenarios where the objective is to select non-repetitive units one at a time, with the option of replacement, from a predefined set of units. Traditional probability calculations in this scenario pose three challenges: the number of floating-point operations to be executed is directly proportional to the chosen set size, susceptibility to floating-point precision errors, and exponential growth in storage needs with increasing number of chosen units. In this scenario, the presented work aims to develop SPM: a sigmoid function-based model that estimates probabilities for such problems with a fixed number of calculations (independent of the input parameter), achieving a constant time complexity algorithm. The research methodology involves generating probability data points, selecting the optimal sigmoid function, augmenting additional data to enhance parameter estimation, identifying parameter estimation equations, and evaluating the model. Moreover, the study's second objective includes training and comparing six established machine learning-based models (including Decision Tree, Random Forest, Support Vector, Linear Regression, Nearest Neighbour, and Artificial Neural Network) against the proposed SPM. The rigorous assessment of the model's performance, utilising metrics including RMSE, MAE and r^2 across a wide range of scenarios involving varying values of the total units, affirms the model's accuracy and resilience. The study findings can improve decision-making processes in various domains, including statistics, cryptography, machine learning and optimisation, by offering a faster, more adaptable solution for probability estimation in units' selection with replacement.

INDEX TERMS Probability estimation, sigmoid function, modeling, non-repetitive units selection, optimization.

I. INTRODUCTION

Probability estimation is a fundamental and indispensable element in various fields, with its significance particularly pronounced when selecting unique units one by one from a predefined set of units, all while allowing the replacement of the selected units in the super-set after each draw. Accurate probability estimations with reduced computation steps are crucial in fields like statistical analysis and decision-making when dealing with complex systems, such as quality control in manufacturing and financial risk assessment [1], [2]. In such a context, the precision of probability estimation plays a pivotal role in facilitating well-informed decision-making processes. Extensive calculations predominantly characterise traditional methods for computing probabilities in this scenario.

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

A general way of calculating the probability in such cases depends on two key parameters: the total number of units in the set, denoted as N and the sample size, represented as n, which indicates the number of units chosen from the set. One of the methods for calculating this probability is based on a series of conditional probabilities, where at each step, you compute the likelihood of selecting a non-repetitive unit and then multiply these probabilities together to get the overall probability [3]. The formula is as follows:

$$P(n,N) = \frac{N}{N} \times \frac{N-1}{N} \times \frac{N-2}{N} \times \ldots \times \frac{N-n+1}{N} \quad (1)$$

Here, the first term, $(\frac{N}{N})$, corresponds to the probability of selecting any unit on the first draw since you have yet to draw any. The second term, $(\frac{N-1}{N})$, represents the probability of choosing a unique (not repeated) unit on the second draw, given that the first unit was unique. The process continues for each subsequent draw, with each term reflecting the probability of selecting a unique unit, considering the uniqueness of the previous draws. This technique involves n multiplications to calculate the probability of selecting unique units with replacement. Furthermore, an alternative method for calculating the probability of selecting unique units from a set with replacement is by using the formula:

$$P(n,N) = \frac{\#favourable\ events}{\#exhaustive\ events} = \frac{{}^{N}P_{n}}{N^{n}} \quad (2)$$

Here, ${}^{N}P_{n}$ represents the number of permutations of choosing n units from N and N^n represents the total number of possible events [4].

Using the traditional formula for probability calculation is susceptible to various errors. Floating-point precision calculation mistakes are a common challenge in numerical computations, leading to errors in various fields. These errors arise due to the finite precision of floating-point representations, resulting in inaccuracies during complex mathematical operations. These mistakes can significantly impact scientific and engineering simulations, emphasizing the need for careful consideration and mitigation strategies when working with floating-point numbers [5]. Consequently, to minimize errors in our study, we opted for calculating probabilities using the second method algorithm (eq. 2), which involves fewer floating-point calculations. However, a limitation of the second method is the exponential growth in storage requirements for large integer numbers as n increases linearly. It is important to highlight that the computational workload (in terms of time-complexity) increases linearly with the size of the selected set, demanding proportional computational resources. In this scenario, our objective is to provide a constant-time complexity algorithm that substantially improves precision without the computational overhead of traditional methods. As the value of N increases, the computational burden escalates. Therefore, establishing a relationship between N, n, and probability is advantageous, enabling a more straightforward probability estimation in extreme cases. To overcome these challenges, the study introduces an innovative application of sigmoid functions for this task.

A sigmoid function is a mathematical function characterized by an S-shaped curve, widely used in various fields including machine learning, statistics, optimization and cryptography [6]. One common example is the logistic function, defined as $f(x) = \frac{1}{1+e^{-x}}$, which maps real numbers to a range of 0 to 1. Sigmoid functions are extensively used in artificial neural networks as activation functions [7]. They exhibit monotonic behaviour and have a bell-shaped first derivative. The logistic sigmoid function, in particular, is invertible and is used in statistics as a cumulative distribution function [8]. These functions are commonly used in logistic regression to model the probability of a binary outcome. Moreover, sigmoid functions are also used in biology and ecology to model population growth and logistic growth [9]. Furthermore, sigmoid functions are part of deep learning models, especially in recurrent neural networks (RNNs) for tasks like natural language processing [10]. These functions are also used in medical diagnosis models to estimate the probability of a patient having a particular condition based on symptoms and test results [11].

The sigmoid functions' flexibility in adapting to different scenarios is a testament to their versatility, making them valuable tools for accurate probability approximation. In our quest to develop a probabilistic estimation formula, we systematically examined probabilities across a range of n:N and total unit counts (N). In the present study, the sigmoid function, encapsulated by the eq. 3 consistently provided a strong fit for these probability patterns.

$$P(x) = \frac{1}{1+e^{\alpha(\beta+x)}} \quad (3)$$

Here, x represents n/N. From further investigation, it was observed that both the parameters α and β displayed varying dependencies on N which can be represented as a similar function, with different set of coefficient values for each parameter. After the model's development, its performance was assessed across a range of N values against the true probability values, utilizing evaluation metrics such as RMSE, MAE, and r^2. Additionally, in the subsequent phase of the study, the model underwent a comprehensive examination by training six machine learning models, namely Decision Tree Regression (DTR), Random Forest Regressin (RFR), Support Vector Regression (SVR), Linear Regression (LR), Nearest Neighbour Regression (NNR), and Artificial Neural Network (ANN). The comparative analysis of these models was conducted on diverse N values, employing metrics like RMSE, MAE, and r^2 for a thorough evaluation. The major contributions of the present study include:

- Introducing a novel approach to probability estimation using sigmoid functions, particularly beneficial in scenarios involving non-repetitive unit selection with replacement.

- Development of a sigmoid-based model with a fixed number of calculations, offering a constant time complexity algorithm for probability estimation.
- Introduction of a novel computational derivation technique to determine the most fitting mathematical equation for the specific problem, enhancing the precision and appropriateness of the developed model.
- Comprehensive comparison of the developed model with widely used machine learning algorithms (DTR, RFR, ANN, LR, SVR, NNR) to establish its superiority in various scenarios.
- Laying the foundation for future studies by exploring adaptability to diverse scenarios and potential applications in various domains, improving decision-making processes in fields such as statistics, cryptography, machine learning, and optimization.

In essence, this research work signifies a promising leap towards creating a faster and more adaptable approach to probability estimation for complex selections. This endeavour presents novel opportunities to enhance decision-making processes in fields where this challenge is inherent, providing advanced solutions for professionals in disciplines such as Engineering, Social Sciences, Finance, Healthcare, Marketing, Agriculture, Education, Quality Control and more.

The outline of the remainder of the article is as follows. Section II provides an in-depth explanation of the methodology employed for developing and evaluating both the intricate and simplified version of the objective sigmoid function. Section III presents the results derived from the simulations to estimate the equation parameters. Section IV discusses the results, their interpretation, associated limitations, implications and prospects for future research. Finally, Section V offers a concise conclusion summarizing the research.

II. METHODOLOGY

The methodology employed in the presented research is structured in two phases, i.e., the SPM development phase and the models' performance assessment phase, which can be divided into seven overall systematic steps (Figure 1).

The initial step systematically generates data points representing the probabilities, characterised by sample and population size variations. Following the generation of data points for diverse N and n:N values, additional data points are required for the fitting of sigmoid functions. To facilitate this, extra probabilistic data points corresponding to n:N values ranging from -0.5 to 1.5 are added. Notably, all probability points corresponding to negative n:N are assigned a value of 1, while those with n:N exceeding 1 are assigned a value of 0. Furthermore, we fit multiple sigmoid-family functions in the probabilistic dataset in the second step, using the Least Square Method (LSM) [12]. LSM is a mathematical technique used to find the best-fitting line or curve through a set of data points. It minimizes the sum of the squared differences between the observed values and the values predicted by the model using eq. 4. Frequently used in linear regression to identify the optimal line portraying the correlation between variables, this method can extend its application to nonlinear models using approaches such as nonlinear least squares.

$$S = \sum_{i=0}^{m} \left(y_i - \hat{y}_i\right)^2 \tag{4}$$

Here, y_i and $\hat{y}_i$ represent the actual and predicted data points, respectively. Next, we determine the equation which best serves the purpose of accurately estimating probabilities. This involved plotting probabilities across a spectrum of n:N ranging from 0 to 1 while varying N within the range of 5 to 1000. Through rigorous analysis, we discerned that the equation with the structure as of eq. 3 consistently provided an ideal fit for the probability plots across different n:N values.

Subsequently, we embark on a parameteric-data collection process. This step entails the acquisition of data relating to parameter values, specifically α and β, across different population sizes (N). Moreover, the step is focused on selecting the most suitable equations for the regression of parameter values concerning their dependence on N. The analysis revealed that the values of both parameters, α and β, can be regressed effectively using the eq. 5 with N as the input variable.

$$f = p_1 \times p_2^{Np_3} \tag{5}$$

In this context, p_1, p_2, and p_3 represent coefficients with distinct values for α and β. After identifying suitable equations, we proceed with the regression process, applying these chosen equations to the parameters and using N as the input variable. In the third step, the research comprehensively evaluates the developed equations. This assessment involved employing well-established statistical metrics, including the Root Mean Square Error (RMSE) [13], eq. 6, Mean Absolute Error (MAE) [13], eq. 7 and the Coefficient of Determination (r^2) [14], eq. 8, across a broad spectrum of scenarios with N values from 50 to 850.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2} \tag{6}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|Y_i - \hat{Y}_i| \tag{7}$$

$$r^2 = \left[\frac{m(\sum Y\hat{Y}) - (\sum Y)(\sum \hat{Y})}{\sqrt{\left[m\sum Y^2 - \left(\sum Y\right)^2\right]\left[m\sum \hat{Y}^2 - \left(\sum \hat{Y}\right)^2\right]}}\right]^2 \tag{8}$$

Here, m is the number of output data point, $\hat{Y}$ is the model estimation and Y is the desired output value. This thorough validation procedure attested to the accuracy and reliability of our formulated equations.

During the study's second phase, we performed a comparative analysis to assess the efficacy of the proposed model

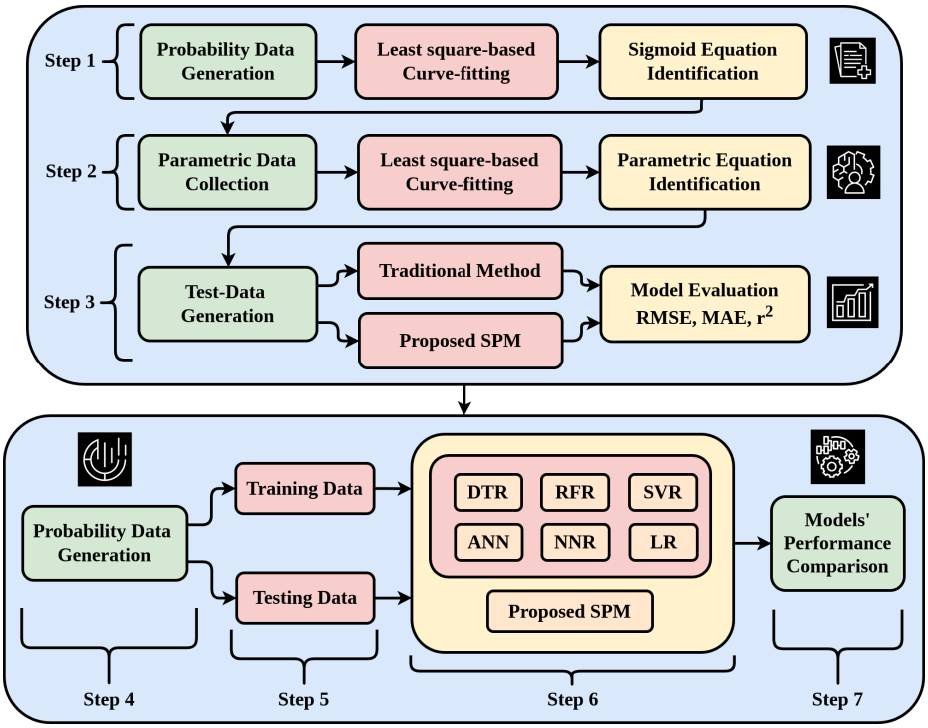


FIGURE 1. Methodological steps undertaken for the development and assessment of SPM.

against established statistical and machine learning models (see Figure 1). To facilitate this evaluation, we partitioned the dataset into two subsets, comprising training and testing data, following an 80:20 ratio.

The practice of dividing data into training and testing sets is fundamental in both machine learning and statistical modelling [15]. The training set is utilized to train models, allowing them to determine patterns and relationships within the data. The independent testing set is a validation tool assessing how well the models generalize to the unseen data. The 80:20 ratio, allocating a larger portion for training, ensures comprehensive model training while maintaining a substantial testing sample for robust evaluation. Additionally, the description of the statistical and machine learning-based models employed in this study's comparative analysis is presented below.

1) Decision Tree Regression (DTR): This approach entails creating a decision tree structure to predict continuous values by segmenting the dataset based on features, allowing it to capture intricate relationships [16]. The importance of each feature i is quantified by the equation presented in 9.

$$M(i) = \sum_{t \in T} \Delta R(i,t), \ \ and$$
$$\Delta R(i,t) = R(t) + R(t_L) + R(t_R), for\ i = 1, 2, \ldots, m \tag{9}$$

Here, $R(t)$ is the sum of the squared deviations at node t for a split i of t into t_L and t_R. The optimal split is the one that maximizes the ΔR. Decision trees excel at capturing non-linear patterns, proving valuable in regression tasks. Through recursive data partitioning, they construct a tree structure featuring decision and leaf nodes, collectively forming an effective predictive model.

2) Random Forest Regression (RFR): RFR is an ensemble machine learning technique that assembles a diverse set of decision trees, collectively known as a "forest." Each tree is created using a random subset of the dataset and a random subset of feature [17]. After K such trees $T(x)_1^K$ are grown, the RFR predictor can be given by eq. 10.

$$\hat{f}_{rf}^K(x) = \frac{1}{K} \sum_{k=1}^{K} T(x) \tag{10}$$

RFR excels in regression tasks by mitigating overfitting and capturing intricate relationships. Its ensemble approach enhances predictive power and generalization, making it a valuable model for diverse applications, including the comparative analysis of probability estimation methods in our study.

3) Nearest Neighbor Regression (NNR): NNR is a machine learning model that estimates values by considering the proximity of data points in the training dataset [18]. It identifies the C closest neighbours (data points) to a given input, where C is a user-defined parameter. The predicted value is computed by averaging the values of these nearest neighbours (eq. 11).

$$\hat{y} = \frac{1}{C} \sum_{c=1}^{C} x_c \tag{11}$$

Here, x_c represents the data point among the closest neighbours of the input value. NNR excels in scenarios where data relationships showcase local patterns or non-linearity, utilizing the similarity of nearby data points to enhance predictive accuracy.

4) Linear Regression (LR): LR is a fundamental statistical approach that models the connection between a dependent variable and independent variables [19]. It aims to determine a linear equation (eq. 12) that optimally fits the data by minimizing the sum of squared differences between observed and predicted values.

$$\hat{y} = \beta x + \epsilon \tag{12}$$

This model offers insights into relationship strength and direction, enabling predictions and variable importance understanding. Widely used in predictive modeling, it serves as a baseline in our study's comparative analysis of probability estimation models.

5) Support Vector Regression (SVR): SVR seeks a hyperplane fitting data with a defined margin for error (epsilon-tube). Its goal is to obtain a function $f(x)$ (eq. (13)) with maximum deviation from training data, yet as flat as possible [20].

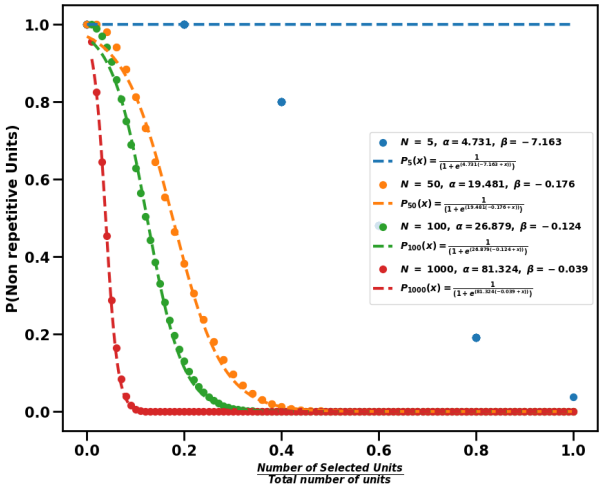
$$f(x) = \sum_{n=1}^{l} w_i K(x) + b \tag{13}$$

Here, l denotes the number of the training data samples, x is the p-dimensional input vector, and K is the kernel function. The model identifies support vector data points closest to the margin, significantly impacting the final prediction. SVR adeptly captures non-linear relationships using kernel functions, transforming the data into a higher-dimensional space.

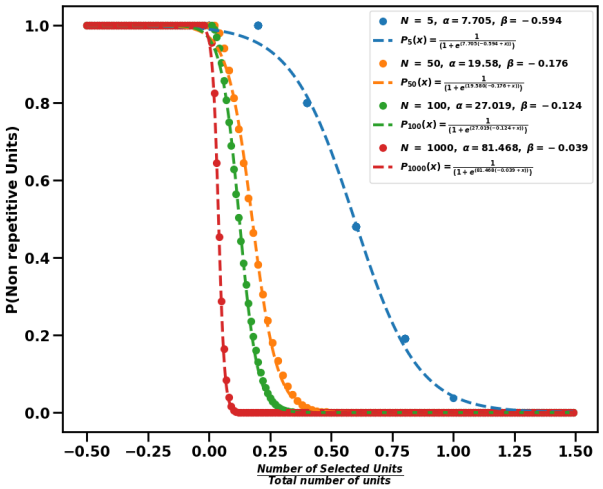
6) Artificial Neural Network (ANN): ANN consists of interconnected perceptron nodes organized in layers: input, hidden, and output. Information travels through the network, and each connection has a weight representing its significance. Every perceptron j present in the network sums its input signals x_i after multiplying them by their respective connection weights w_{ji}. It then applies an activation function to the resultant and passes the output to the next layer. The working of a perceptron can be mathematically described by eq. (14).

$$y_j = \psi\left(\sum_{i=1}^{u} w_{ji} x_i\right) \tag{14}$$

where ψ is the activation function utilizing the weighted summations of the inputs, and u represents the number of nodes in the previous layer. Some of the common activation functions used by the model developers are the sigmoid function, hyperbolic tangent function [21] and ReLU [22]. Learning occurs through an iterative process called backpropagation, where the model adjusts weights to minimize the difference between predicted and actual outcomes. The hidden layers allow ANNs to model complex relationships and patterns within the data. This architecture enables ANNs to adapt and generalize well, making them powerful tools in various machine-learning applications.



(a)



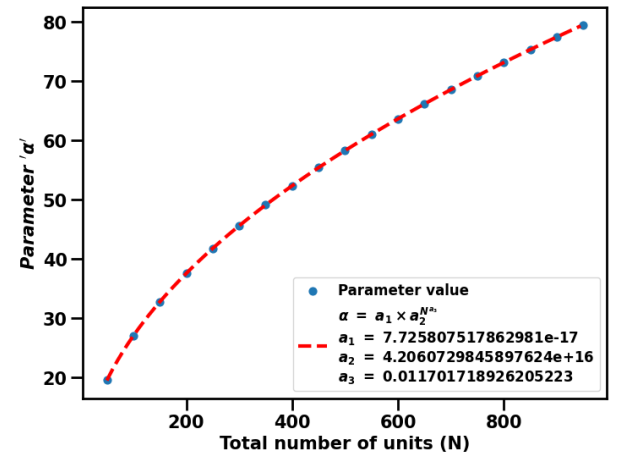
(b)

FIGURE 2. Plot of the probability values calculated using eq. 2 along with the regressed line using eq. 3; (a) without appending additional data points, (b) after appending additional probability datapoints.

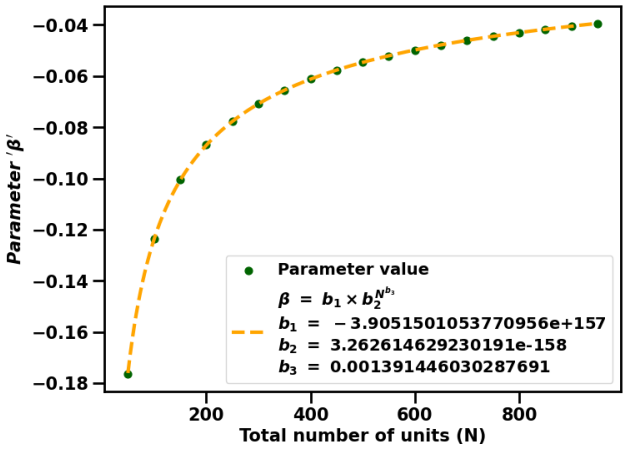
We employed a distinct testing dataset to measure models' performance on unseen data, evaluating both established models and the proposed SPM. Input parameters N and n:N represented test scenario characteristics with N ranging from 50 to 1000. Moreover, the estimation power of the models was assessed using RMSE, MAE, and r^2. This comparative analysis affirmed the models' effectiveness and suitability for probability estimation in diverse real-world scenarios.

III. EXPERIMENTS AND RESULTS

Figure 2 (a) depicts the probabilities corresponding to varying n:N ratios with N values of 5, 50, 100, and 1000. Additionally, the figure displays the regressed sigmoid line



(a)



(b)

FIGURE 3. Regressing the values of (a) α using eq. 15 and (b) β using eq. 16 with N as input.

on the probability data points. From the figure, it is observed that the sigmoid curves exhibit steeper slopes with increasing N. Furthermore, the plots demonstrate that the identified sigmoid function (eq. 3) accurately fits the probability datapoints, verifying its efficacy for probability estimation in our context. Notably, the sigmoid function for N = 5 faces fitting challenges due to insufficient data points. To address this, we have augmented the dataset with additional probability data points, extending the x-axis from -0.5 to $+1.5$, as illustrated in figure 2 (b). The extended data points enhance the fitting of the sigmoid function to the probability points.

After determining the most effective sigmoid function for the given probability problem, the data corresponding to the parameters α and β of the identified function, along with their respective N values, was gathered. eq. 3 encompasses these two parameters. Subsequently, we utilized this parametric dataset to ascertain optimal mathematical expressions describing the relationship between these parameters and N. Through experimentation, we determined that eq. 15 provides the most accurate fit for α (Figure 3 (a)):

$$\alpha = a_1 \times a_2^{N^{a_3}} \tag{15}$$

Further investigation revealed that the same equation could be employed to regress the β parameter (eq. 16) using a distinct set of coefficient values (Figure 3 (b)). Table 1 presents the numerically calculated coefficients for equations 15 and 16.

$$\beta = b_1 \times b_2^{N^{b_3}} \tag{16}$$

TABLE 1. Coefficient values calculated for the identified parameters equations.

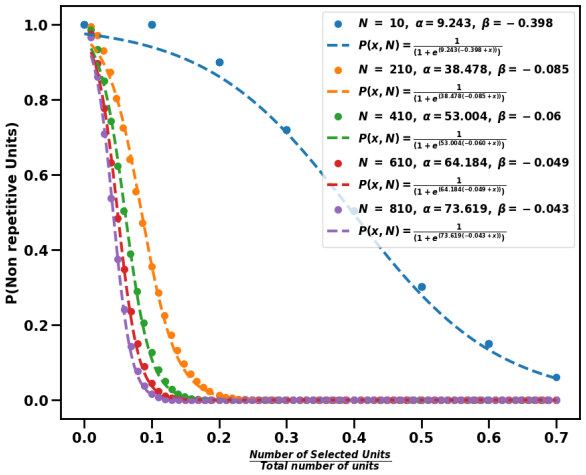
S.No.	Coefficient	Value
1.	a_1	7.725807517862981E-17
2.	a_2	4.2060729845897624E+16
3.	a_3	0.011701718926205223
4.	b_1	3.9051501053770956E+157
5.	b_2	3.262614629230191E-158
6.	b_3	0.001391446030287691

Following the development of the proposed SPM, the study conducted a comprehensive performance assessment across a diverse range of N values from 50 to 850. Figure 4 (a) visually compares actual and estimated probability values across varying N, showcasing the model's robust performance. Subsequently, Figure 4 (b) quantitatively evaluates the model's performance using metrics such as RMSE, MAE, and r^2. Notably, the figures illustrate that the model's RMSE and MAE decrease as N increases, indicating enhanced accuracy with a larger total number of units. The RMSE ranges from 0.006200 to 0.013551, and the MAE ranges from 0.001519 to 0.007009. This trend is consistent with r^2, where a larger N corresponds to better goodness of fit, with r^2 ranging from 0.998161 to 0.998438.

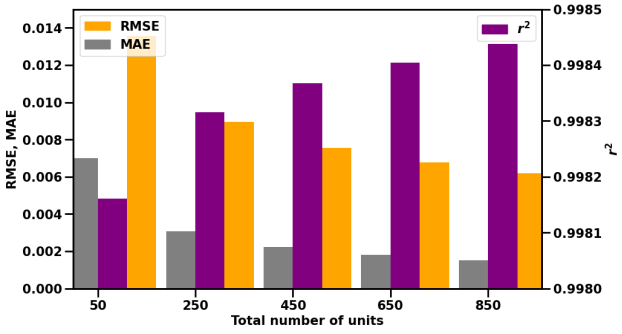
Figure 5 presents various models' performance metrics (RMSE, MAE, and r^2), including the SPM, DTR, RFR, ANN, LR, SVR and NNR. The analysis shows that the SPM exhibits excellent performance with a low RMSE and MAE, indicating accurate predictions (table 2). The high r^2 value (close to 1) suggests a strong fit, confirming the model's effectiveness in capturing the relationship between variables. Moreover, RFR, DTR and ANN perform well with relatively low RMSE and MAE, indicating good accuracy. The high r^2 value suggests a strong predictive capability, although slightly lower than the SPM. Furthermore, SVR, LR and NNR show comparatively higher errors (RMSE and MAE) and a lower r^2 value, suggesting that it may struggle to capture the complexity of the underlying relationships in this context. The analysis indicates that the proposed SPM outperforms other models, demonstrating its effectiveness in probability estimation for non-repetitive unit selection. The Tree-based models (RFR, DTR) generally perform well, while SVR, NNR, and LR show comparatively lower accuracy.

TABLE 2. Preformance comparison of the proposed SPM with six ML-based regression models.

S.No.	Model	RMSE	MAE	r^2
1.	SPM	0.01131414475	0.00379360614	0.9982202167
2.	RFR	0.02880895791	0.008985694208	0.9884607188
3.	DTR	0.03517940801	0.008340815094	0.9827931762
4.	ANN	0.0369162548	0.02359100008	0.9810521941
5.	SVR	0.2119314272	0.08495453118	0.3755259671
6.	NNR	0.223529966	0.09274821238	0.3053034165
7.	LR	0.2485444213	0.1509232229	0.1411215658



(a)



(b)

FIGURE 4. Visualising (a) the probability values calculated using eq. 2 in comparison with the values calculated using the proposed SPM and (b) RMSE, MAE and r^2 of the SPM over varying N from 50 to 850.

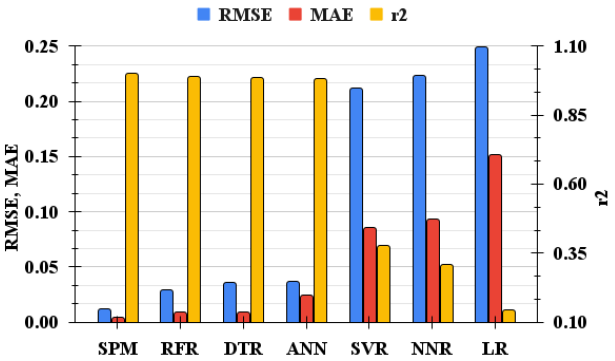


FIGURE 5. RMSE, MAE and r^2-based comparison of the proposed SPM and six ML regression models.

IV. DISCUSSION

The obtained results in this study are noteworthy and provide valuable insights into probability estimation using sigmoid equations. The results demonstrate that the sigmoid eq. 17 effectively captures the relationship between sample size-to-population size ratio and probability.

$$P(x) = \frac{1}{1 + e^{\alpha(\beta+x)}},$$
$$\alpha = a_1 \times a_2^{N^{a_3}}, \text{and}$$
$$\beta = b_1 \times b_2^{N^{b_3}} \tag{17}$$

With the coefficient values as given in table 1, this equation offers a practical and accurate approach for estimating probabilities, especially in scenarios involving selecting non-repetitive items from a set with replacement. The observed variations in the α and β parameters with changing population size highlight the complexity of the relationship. The ability to model these parameter-changes provides a tailored approach to probability estimation, adapting to different contexts. This is a unique and valuable insight into the behaviour of the probability function. The results showcase the models' accuracy across a broad range of population sizes, from 50 to 850. This adaptability underscores the versatility of the sigmoid equation in providing precise probability estimations in various scenarios.

The strong fit of the model to the data is evident in the high r^2 values exceeding 0.99. These high r^2 scores affirm the reliability and goodness of fit of the proposed model across different population sizes, reinforcing its efficacy. The decreasing RMSE values as population size increases indicate improved accuracy in probability estimation with larger datasets. This observation highlights the model's ability to handle increasingly complex scenarios. Nonetheless, it is to be noted that the simulation results should be interpreted with caution, as they are subject to precision errors inherent in floating-point arithmetic [5]. To address this limitation, future research endeavors will focus on substantiating this approximation through a derivational approach as an alternative to the simulation-based method.

The proposed sigmoid-based model offers a streamlined approach to estimating probabilities, reducing the computational burden associated with traditional methods requiring numerous calculations. The established methodologies are characterized by a time complexity of $O(n)$, i.e. the number of calculations are dependent on the number of chosen units, in contrast to the proposed approach which offers a remarkable improvement, enabling the calculation of the desired probability in constant time, denoted as $O(1)$ [23].

In the study's second phase, a comparative analysis was conducted to assess the efficacy of the SPM against existing models. The dataset was partitioned into training and testing subsets, and model performances were evaluated using RMSE, MAE, and r^2. The results indicate that the DTR, RFR and ANN models exhibit the lowest RMSE values, signifying precise predictions among existing models. Conversely, LR, NNR, and SVR models show higher RMSE values, indicating less accuracy in predictions. Similar trends were observed in terms of MAE and r^2. Notably, our findings align with previous studies where RFR and DTR consistently outperformed other ML algorithms.

The study conducted by [24] systematically compared three machine learning techniques, namely Gaussian process regression, RFR, and SVR, within a GEOBIA framework. The results revealed that RFR outperformed conventional regression by 48%, demonstrating superior burn severity assessment and reduced sensitivity to variations in remote sensing variable combinations. In the research conducted by [25], amylase and urease activities were estimated using Multiple Linear Regression (MLR) and RFR models with various covariates. The RFR model exhibited superior performance over MLR, attributed to its capacity to handle nonlinear relationships and hierarchical dependencies, resulting in lower errors and enhanced accuracy.

Another study by [26] focused on landslide susceptibility mapping and compared logistic regression with RFR, utilizing hyperparameter optimization through the Bayesian algorithm. The RFR model demonstrated superior stability and predictive capability in this assessment. In the analysis conducted by [27] on COVID-19 cases in Indonesia, the study predicted new cases using DTR and LR algorithms. The DTR algorithm achieved higher r^2 scores (95.69% for training and 92.15% for testing) compared to LR (79.93% for training and 77.25% for testing).

Furthermore, the pivotal outcomes of this study reveal that the proposed Sigmoid-based Probability Estimation Model (SPM) exhibited superior performance compared to alternative models, as evidenced by the lowest RMSE and MAE values. These metrics underscore the SPM's heightened accuracy in estimating probabilities. Notably, the inverse relationship between RMSE, MAE, and the population size (N) signifies the SPM's enhanced precision with larger datasets. Additionally, the SPM demonstrated a notably higher coefficient of determination (r^2) compared to other models when evaluated on the testing dataset. This emphasizes the SPM's robust capability to elucidate the variance observed in probability data, further affirming its efficacy in probability estimation.

While our study has made substantial strides in probability estimation, it is crucial to acknowledge a limitation, namely, the primary focus on probability estimation in scenarios of one-by-one units' selections with replacement. The generalizability of our findings to diverse probability estimation problems is constrained, as each problem may present unique challenges and require tailored solutions. The research outlined in this study lays a robust foundation for future inquiries and advancements in probability estimation. Subsequent investigations could extend the model's applicability to a broader spectrum of probability estimation problems, encompassing those with intricate sampling methods or dependencies between selections.

Moreover, although we assessed traditional ML models, there exists an opportunity to explore more advanced techniques such as Deep Learning and ensemble methods for probability estimation. These advanced methods might offer enhanced accuracy and adaptability across diverse scenarios. Additionally, exploring the synergy of different probability estimation models, including the proposed SPM and ML algorithms, to create hybrid models leveraging the strengths of each approach could lead to optimized solutions for specific use cases. Furthermore, investigating approaches to make probability estimation viable in resource-constrained environments, where computational resources are limited, is imperative. This could involve the development of lightweight models or algorithms tailored to operate efficiently in such environments.

V. CONCLUSION

In summary, the presented research tackled the problem of estimating probabilities when drawing units one at a time from a set without repeating any. The conventional methods for calculating probability in such scenarios entail high computational expenses in terms of time and space complexity. The study introduced a novel SPM model based on sigmoid functions that simplify the estimation process using eq. $P(x) = \frac{1}{1+e^{\alpha(\beta+x)}}$. Furthermore, the regression of parameters α and β against the variable N successfully mitigated the challenge of equation variations across different N values. The study followed a methodical approach involving several steps: data generation, equation assessment, parameter selection, data collection, regression using the least squares method, equation evaluation and model assessment using RMSE, MAE and r^2. The study results showed that the proposed model provides accurate probability estimations with reduced computational effort. Furthermore, a comparative analysis involving the proposed SPM and six ML-based models (DTR, RFR, ANN, LR, SVR and NNR) revealed that the SPM exhibited superior performance, surpassing all other models. The model's strong fit to data across various population sizes highlights its reliability. The implications of this research are far-reaching, offering practical benefits in fields demanding

precise probability estimations, such as statistics and optimisation. The future scope of this research involves exploring the applicability of the sigmoid-based model in various real-world scenarios and extending it to address more complex probability estimation challenges. Overall, this work advances probability estimation methodologies and opens doors for enhanced decision-making processes across diverse fields.

ACKNOWLEDGMENT

The authors would like to thank Yash Arora and Sumit Mahajan, Ph.D. scholars with the Department of Mathematics, IIT Roorkee, for their invaluable contributions, insightful discussions, and constructive ideas while developing this work.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The dataset used in the current research study is publicly available.

REFERENCES

[1] T. Parhizkar, J. E. Vinnem, I. B. Utne, and A. Mosleh, "Supervised dynamic probabilistic risk assessment of complex systems—Part 1: General overview," *Rel. Eng. Syst. Saf.*, vol. 208, Apr. 2021, Art. no. 107406.
[2] J. V. B. de la Paz, L. A. Rodríguez-Picón, V. Morales-Rocha, and S. V. Torres-Argüelles, "A systematic review of risk management methodologies for complex organizations in Industry 4.0 and 5.0," *Systems*, vol. 11, no. 5, p. 218, Apr. 2023.
[3] J. K. Blitzstein and J. Hwang, *Introduction to Probability*. Boca Raton, FL, USA: CRC Press, 2019.
[4] H. Le Bras, "Feller William—An introduction to probability theory and its applications," *Population*, vol. 23, no. 2, p. 375, 1968.
[5] D. Goldberg, "What every computer scientist should know about floating-point arithmetic," *ACM Comput. Surv.*, vol. 23, no. 1, pp. 5–48, Mar. 1991.
[6] N. Kyurkchiev, A. Iliev, and A. Rahnev, *Some Families of Sigmoid Functions: Applications to Growth Theory*. Saarbrücken, Germany: Lap Lambert Academic, 2019.
[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[8] N. Kyurkchiev and S. Markov, *Sigmoid Functions: Some Approximation and Modelling Aspects*. Saarbrücken, Germany: Lap Lambert Academic, 2015.
[9] D. P. Smith, N. Keyfitz, and P.-F. Verhulst, "A note on the law of population growth," in *Mathematical Demography*, 1977, pp. 333–339. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-81046-6_37
[10] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[11] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. London, U.K.: Oxford Univ. Press, 2003.
[12] Å. Björck, "Least squares methods," in *Handbook of Numerical Analysis*, vol. 1. Amsterdam, The Netherlands: Elsevier, 1990, pp. 465–652.
[13] T. Chai and R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)," *Geosci. Model Develop. Discuss.*, vol. 7, no. 1, pp. 1525–1534, 2014.
[14] I. S. Helland, "On the interpretation and use of R^2 in regression analysis," *Biometrics*, vol. 43, no. 1, pp. 61–69, Mar. 1987.
[15] R. R. Picard and K. N. Berk, "Data splitting," *Amer. Statistician*, vol. 44, no. 2, pp. 140–147, 1990.
[16] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data," *Remote Sens. Environ.*, vol. 97, no. 3, pp. 322–336, 2005.
[17] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, Dec. 2015.
[18] N. S. Altman, "An introduction to Kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.
[19] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 4, no. 3, pp. 275–294, 2012.
[20] M. Awad, R. Khanna, M. Awad, and R. Khanna, "Support vector regression," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA, USA: Apress, 2015, pp. 67–80.
[21] I. S. Isa, Z. Saad, S. Omar, M. K. Osman, K. A. Ahmad, and H. A. M. Sakim, "Suitable MLP network activation functions for breast cancer and thyroid disease detection," in *Proc. 2nd Int. Conf. Comput. Intell., Modeling Simulation*, Sep. 2010, pp. 39–44.
[22] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with ReLU activation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 597–607.
[23] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2022.
[24] C. Hultquist, G. Chen, and K. Zhao, "A comparison of Gaussian process regression, random forests and support vector regression for burn severity assessment in diseased forests," *Remote Sens. Lett.*, vol. 5, no. 8, pp. 723–732, Aug. 2014.
[25] X. Xie, T. Wu, M. Zhu, G. Jiang, Y. Xu, X. Wang, and L. Pu, "Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land," *Ecological Indicators*, vol. 120, Jan. 2021, Art. no. 106925.
[26] D. Sun, J. Xu, H. Wen, and D. Wang, "Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest," *Eng. Geol.*, vol. 281, Feb. 2021, Art. no. 105972.
[27] D. Darwin, D. Christian, W. Chandra, and M. Nababan, "Comparison of decision tree and linear regression algorithms in the case of spread prediction of COVID-19 in Indonesia," *J. Comput. Netw., Archit. High Perform. Comput.*, vol. 4, no. 1, pp. 1–12, Jan. 2022.

SAMARTH GODARA received the M.Tech. degree from the National Institute of Technology, Jalandhar, Punjab, India, and the Ph.D. degree from IIT Roorkee, specializing in artificial intelligence in agriculture. He contributes to the scientific community as a Scientist with the ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. His research interests include artificial intelligence, machine learning, and data analytics. He has authored numerous research articles in big data analytics and deep learning.

G. AVINASH received the M.Sc. degree in agricultural statistics from the University of Agricultural Sciences, GKVK, Bengaluru, India, making him a prominent figure in advanced statistical analysis and data analysis. He is currently an accomplished Researcher in statistical analysis with the ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. He has authored numerous high-impact research articles in the fields of statistical analysis, deep learning, and time series forecasting, which have been published in prestigious international journals.

RAJENDER PARSAD received the Ph.D. degree in agricultural statistics. He is currently a highly accomplished Fellow with a profound academic journey culminating in the Ph.D. degree. He is also the Director of the ICAR-Indian Agricultural Statistics Research Institute, New Delhi, he leads with distinction. He was honored with the NAAS Recognition Award, in 2015 and 2016, for his significant contributions to the field of social sciences. In addition, he received the National Award in Statistics, in 2010 to 2011, for his outstanding work in statistics, the Prof. PV Sukhatme Gold Medal Award, in 2010, from the Indian Society of Agricultural Statistics, and he held the ICAR National Fellow title, from January 2005 to April 2009, recognizing his remarkable achievements in the field of agriculture and statistics. He is renowned for his contributions to agricultural statistics, design of experiments, sampling techniques, and statistical computing. His extensive leadership roles, awards, and fellowships underscore his invaluable impact on statistics and agriculture.

SUDEEP MARWAHA received the B.Sc. degree in electronics from the University of Delhi, New Delhi, India, in 1995, the M.Sc. degree in computer applications from the Indian Agricultural Research Institute (IASRI), New Delhi, and the Ph.D. degree in computer science from the University of Delhi, in 2008. He is currently a Principal Scientist and a Professor with the Division of Computer Applications, ICAR-IASRI. He completed the following projects, including solution architect for semantic web-enabled systems, management information systems, ERP (Oracle Apps), knowledge base systems, and image analysis-based systems. He has published more than 50 research articles. His research interests include artificial intelligence, semantic web, and ontologies.

MUKHTAR AHMAD FAIZ is currently a Lecturer with Kandahar University, Afghanistan. He is also a Distinguished Scholar with a Doctor of Philosophy with the Indian Agricultural Research Institute, specializing in agronomy. His expertise encompasses sustainable agriculture, crop production, soil fertility, plant nutrition, and agricultural development. With a focus on organic farming, climate change impact on agriculture, and soil conservation, he has contributed significantly to nutrient management, precision agriculture, and food security. His multidisciplinary skills extend to agroecology, field experimentation, and the promotion of environmentally sustainable practices in agriculture.

RAM SWAROOP BANA received the M.Sc. and Ph.D. degrees from the esteemed Indian Agricultural Research Institute (IARI), New Delhi, India. He is currently a distinguished Senior Scientist with a wealth of expertise in the field of agronomy and vegetable science. His dedication to the study of agriculture has made him an invaluable contributor to the institute's mission. With an extensive background in agronomy, he brings a wealth of knowledge and experience to the research community. His distinguished academic background is complemented by a prolific research portfolio, encompassing over 50 research articles published in renowned international journals.