International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 43 No. 1 Jul. 2024, pp. 198-206

http://www.ijias.issr-journals.org/

A predictive model for 400-metre performance: Analysis using the first 100 m and athlete age

Papa Serigne Diène[1], El Hadji Mama Guène[2], Daouda Diouf[1], Ndiack Thiaw[1], Mame Ngoné Bèye[1], Ndarao Mbengue[1], Abdoulaye Samb[3], and Abdoulaye BA[3]

[1]STAPS-JL Laboratory, INSEPS-UCAD, Senegal

[2]Department, LGL-TPE, University of Lyon 1, Lyon, France

[3]Faculty of Medicine and Pharmacy Odontostomatology, UCAD, Senegal

ABSTRACT: In this study, a linear regression approach is used to model 400 m performance. We have choosen to consider the time of the first 100 metres and the age of the athletes as key variables, as they are likely to play a determining role in succeeding in this specific distance.
The start, symbolised by the first 100 metres, is often considered a crucial phase in the 400m. Sprinters who manage to maintain rapid acceleration in this first part of the race tend to perform more consistently over the whole 400 metres.
Studies have shown that competitive experience can play a significant role in athletic performance. Athletes who have accumulated years of experience often develop more efficient running strategies and better effort management, thus positively influencing their results.

KEYWORDS: machine Learning, linear regression, 400m race, modelling.

1 INTRODUCTION

The 400 metres, often referred to as the "queen race", represents a unique challenge in athletics. This distance requires a subtle balance between the explosive speed of the sprint and the endurance needed to maintain the pace throughout the race.

Modelling 400m performance is crucial to optimising athlete training. Understanding the key factors that influence success in this discipline means that training programmes can be adapted more precisely, offering a competitive advantage.

Previous studies have explored various models for predicting athletic performance. Thomas R et al (1989), have produced various descriptive models that trace all the main factors that impact on the athlete's final result [1]. Alderman (1974) already distinguished four families of sports performance factors: motor abilities, technical skills, physical characteristics and psychological and behavioural elements [2]. As for Carron (1980), he extended Cratty's (1967) model [3] still with four groups of factors: social factors, structural factors, physiological factors and psychological factors [4].

In the same vein, Weineck (1983) proposes a model based on four determinants: constitutional factors, physical abilities, technical-tactical skills and personality factors [5]. Franks and Goodman

(1986) adapted the model of Calvert *et al* (1976) [6] to propose a model highlighting the multiple interactions within a complex system involving several variables including: physical abilities, technical skills, motivation and emotional factors [7].

In sports performance modelling, various machine learning approaches are applied. According to a research study published in the Journal of Sports Sciences (Baca & Dabnichki, 2020), these approaches include neural networks, classification and regression methods, deep-learning, as well as natural language processing techniques for analysing sports texts and

unstructured data such as match commentaries. These methods are used to predict athlete performance, optimise training and game strategies, and provide real-time predictive analysis during competitions.

Given that it is impossible to maintain the highest speed from the beginning to the end of the race, it will necessarily be necessary to find the most ergonomic way of distributing speed and energy over the total distance. Arnold (1989) states that great 400 runners generally have one (1) second difference between their best performance in the 200m and the time taken to pass the first 200m in the 400m [8]. This is how Gambetta (1978) suggests that a sense of the right distribution of pace and effort is, therefore, a must [9]. He considers that the first runner to go under 44 seconds (Lee Evans, USA) completed the first 200m in 21.2 seconds, i.e. 0.5 seconds off his best performance over 200m, to achieve a performance of 43.86 seconds. This explains why the best athletes have a smaller differential.

Senegalese athletes have often succeeded in this speciality. Amadou Dia BA won Senegal's only Olympic medal at the 1988 Olympic Games in Seoul, with a performance of 47'23 in the 400m hurdles. Amy Mbacké THIAM won gold in the women's 400m at the 2001 World Championships in Edmonton with a time of 49'86, which is also the current Senegalese record. It wasn't until 2022 that we recorded three 400m runners under 47".

It would seem appropriate to model the 400m using the best Senegalese in order to make these results sustainable.

In this study, we adopt a linear regression approach to model 400 m performance. The start, symbolised by the first 100 metres, is often considered to be a crucial phase in the 400 m. Studies have shown that age can play a significant role in athletic performance.

The aim of our investigation is to use the time achieved in the first 100 m to create a mathematical model that can be used to predict performance in the 400 m race.

Once achieved, this objective will contribute to a better understanding and optimisation of athletic performance in general and 400-metre performance in particular, with potential implications for sports training and athlete preparation.

2 METHODOLOGY

2.1 EQUIPMENT

We carried out our study at the maître Abdoulaye WADE annex stadium and at the STAPSJL laboratory of the INSEPS located within the Stade Iba Mar DIOP.

The study involved a sample of 18 Senegalese athletes taking part in the national championships, who gave their informed written consent in accordance with the inclusion criteria requiring regular participation in competitions organised by the Senegalese Athletics Federation and the achievement of a performance between 45.48 and 51.99. Athletes who were not 400m specialists and those who competed outside Senegal were excluded from the study.

We carried out our experiment during the Senegal 2023 national athletics championships, with a data set including 400m performance, split 100m times for each 100m of the 400m competition, and relevant variables such as age, gender, height, weight, etc. The data set was then compared with the data from the Senegal 2023 national athletics championships.

For each athlete, we collected the 400m performance and the time taken to run each 100m sequence during the race.

The athlete's performance at the end of the 400m is immediately recorded on the display board of the Time tronics Argus electronic stopwatch.

As regards the times taken to cover each 100 m sequence during the race, we marked out the start and finish of each 100 m sequence visibly on the track (fluorescent markers). The video of each race was taken from four angles using 4 SONY compact professional Full HD NXCAM cameras and analysed using DARTFISH image processing software to determine the time taken in each 100m sequence.

We then took each athlete's date of birth from their licence to determine their chronological age. Chronological age is the number of days and years since birth.

2.2 METHODS

2.2.1 DESCRIPTION OF LINEAR REGRESSION AS A MACHINE LEARNING TECHNIQUE

Linear regression is a machine learning technique designed to model a target variable using one or more independent variables. Its objective is to establish a linear relationship between a dependent variable Y and several variables X, using the mathematical formalism:

$Y= \beta0+\beta1X1+\beta2X2+...+\beta nXn +\epsilon$ Where $\beta0, \beta1, \beta2,..., \beta n$ represent the parameters of the model to be determined, Y is the target variable X1, X2,..., Xn are the independent variables, and ϵ is the residual error that is not explained by the model.

The variable Y and the variables X are said to be linearly correlated. Correlation measures the relationship between two variables. If a correlation between two variables is detected, a Student's t test is required to assess its significance. Linear correlation, also known as Pearson correlation, quantifies the strength of the linear relationship between the two characteristics. It varies from -1 to 1. A negative correlation between X and Y indicates that a small increase in X leads to a decrease in Y, and vice versa. A positive correlation between X and Y means that both variables increase in the same direction.

The term $\beta0+\beta1X1+\beta2X2+...+\beta nXn$ represents the model's prediction, noted $\hat{Y}$, and so the residual error ϵ is defined as the difference between the observed value Y and the prediction Y, i.e. $Y- \epsilon = Y- \hat{Y}$. The aim is to minimise this residual error ϵ, ideally until it is equal to zero if the model is perfect.

An effective approach to finding the best parameters is to use the gradient descent method by minimising the cost function, which is usually defined as the mean square error ϵ^2. This method consists of initializing the parameters with random values, then adjusting these values at each iteration in order to minimize ϵ^2.

At each iteration of the gradient descent, the values of the parameters are adjusted in proportion to the gradient of the cost function with respect to these parameters. This iterative process is repeated until a stopping criterion is reached, such as a predefined number of iterations or when the decrease in the cost function becomes negligible.

2.2.2 CHOICE OF INDEPENDENT AND DEPENDENT VARIABLES

The choice of variables in our linear regression model to model the 400m in athletics is crucial to ensure the relevance and accuracy of the predictions. The independent variables are those that influence the dependent variable, which we are trying to predict or explain.

The correlation results show distinct relationships between the variables studied. A very strong correlation (0.95) between the 400m and 100m times suggests a strong dependency between these two variables. Thus, the 100m time is considered to be an independent variable, strongly influencing the total 400m time, which is the dependent variable.

Similarly, a moderate correlation (-0.53) between 400m time and age indicates an inverse relationship: older athletes tend to have faster 400m times. This suggests that age is also an independent variable influencing 400m performance, although its impact is less pronounced than that of 100m time.

The selection of the time in the first 100m as an independent variable is based on its crucial role in the dynamics of the 400m race. It reflects the ability to establish a rhythm and generate power from the start of the race. The inclusion of age as an independent variable is based on the hypothesis that experience and physical maturity evolve with age, thus influencing performance.

By combining these variables in our model, we aim to develop a robust method for predicting and understanding performances over the 400m in athletics. This approach will enable an in-depth analysis of the factors that influence athletes' performances over this iconic distance, thereby helping to improve training and competition strategies.

2.2.3 DATA COLLECTION AND PRE-PROCESSING

Data pre-processing is a crucial phase before modelling begins. This stage has several components, depending on the structure of the data available, such as detecting outliers, removing or filling in missing values, and normalising the data. The main objective is twofold: to format the data in the correct format and to improve the performance of the machine learning models.

In our study, the data are standardised, which means that all the variables are transformed so that they have a mean of zero and a standard deviation of 1. These transformed variables are referred to as reduced variables. The transformation is performed as follows: $Xr= (X-\mu) /\sigma$.

Standardisation aims to standardise the scale of all the variables, ensuring that they have equivalent importance in Machine Learning algorithms, some of which are sensitive to extreme values in the data. The data are then randomly mixed and divided into two distinct sets:

- A training set, designed to train the model and determine its parameters. The Root Mean Square Error (RMSE) and the score are calculated. This set generally contains between 75% and 80% of the data
- A test set, comprising between 25% and 20% of the total data. These data represent new observations that the model has never encountered. The RMSE is also calculated for these data, as is the score. The RMSE of the test data should normally be higher than that of the training data, as the model generally makes more errors on new observations. As a result, the prediction score for the test data is also affected

2.2.4 APPLICATION OF LINEAR REGRESSION TO MODEL 400-METRE PERFORMANCE

Among the various variables available, we selected the time achieved over the first 100 metres (noted T100) and the athlete's experience in the 400 metres (Exp) as independent variables, because they correlate with performance in the 400 metres (Perf).

We chose to use linear regression to model the relationship between performance (Perf) and these independent variables, namely T100 and Exp. Initially, we simplified the model by considering that performance depends solely on the time achieved over the first 100 metres. The associated linear regression equation is as follows: $Perf = \beta0 + \beta1 \times T100 + \epsilon$, where $\beta0$ and $\beta1$ represent the coefficients of the model, reflecting the impact of the time achieved over the first 100 m on performance, while ϵ denotes the error term.

To improve the performance of the model, we decided to incorporate the athlete's experience in running the 400m, as we believe that this experience can make a difference between runners, particularly with regard to effort management during the race. Thus, the linear regression equation evolves to include this additional variable: $Perf = \beta0 + \beta1 \times T100 + \beta2 \times Exp + \epsilon$.

The model coefficients are estimated using the least squares method, using the gradient descent algorithm in both cases. This step, known as model training, is crucial for obtaining accurate estimates of the coefficients.

Finally, to assess the performance of the regression, part of the model is tested on new data that the model has never encountered. This validates the model's ability to generalise and accurately predict 400- metre performance on new observations.

3 RESULTS

3.1 UNIVARIATE ANALYSIS OF THE DIFFERENT VARIABLES USED TO MODEL THE 400M BEFORE THE RACE

Table 1. Presentation of the data obtained relating to chronological age, training age, specific training age for the 400m and performance achieved during the 400m competition

N= 18	PERF in tests (seconds)	Best 400m performance (seconds)	Age (year)	Training age (year)	Training age 400m (year)
Average	50,07	48,87	22,56	4,11	3,78
Standard deviation	±1,78	±1,49	±3,03	±2,27	±2,10

Table 2. Times achieved by athletes in each 100m (100m split) during the 400m race

N=18	400m sequences: times achieved in each 100m sequence)			
	Time 1er 100m (s)	Time 2e 100m (s)	Time 3e 100m(s)	Time 4e 100m(s)
Average	12,44	12,50	12,69	12,45
Standard deviation	±0,49	±0,57	±0,64	±0,52

Table 3. Individual and average values for frequency, amplitude and number of strides for each 100m split in the 400m competition

N°=18	1er 100m			2e 100m			3e 100m			4e 100m		
	Strides per 1er 100m	Stride frequency at 1er 100m (Foulé/sec)	Stride amplitude at 1er 100m (metre)	No. of strides at 2e 100m	Stride frequency at 2e 100m (Strides/sec)	Stride amplitude at 2e 100m (metres)	No. of strides at 3e 100m	Stride frequency at 3e 100m (Strides/sec)	Stride amplitude at 3e 100m (metre)	Number of strides at 4e 100m	Stride frequency at 4e 100m (Strides/sec)	Stride amplitude at 4e 100m (metre)
Average	48,78	0,26	2,05	48,33	0,26	2,07	50,72	0,26	2,03	47,94	0,26	2,09
Standard deviation	±1,77	±0,01	±0,07	±2,52	±0,01	±0,10	±11,42	±0,03	±0,27	±2,01	±0,01	±0,08

Table 4. Individual and average lactatemia values of athletes at rest, after warm-up and at the end of the 400m race

N=18	Rest	After warming up	After 400m
Average	1,32	2,23	18,55
Standard deviation	±0,30	±0,32	±2,56

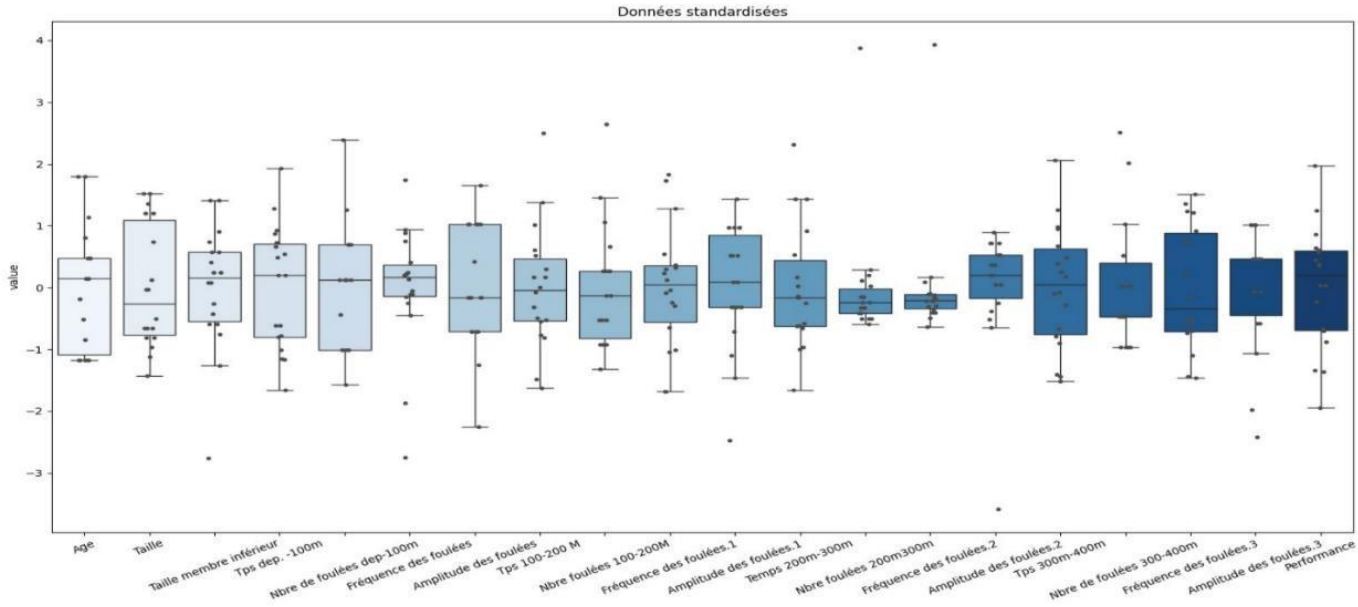

Fig. 1. Illustration of the dispersion of variables around the median

Our boxplot shows the dispersion of subjects around the median for each variable studied. Sometimes there are outliers, but they are generally located on variables with a non-significant correlation with the 400m time. For example, the number of strides in the first 100 m of the race (r: 0.45) showed three (3) data points outside our boxplot. However, it should be noted that most of the data are found in the boxplot, mainly in the second and third quartiles.

The standardisation of the data allows us to have a better representation of the variables through the moustache box below:

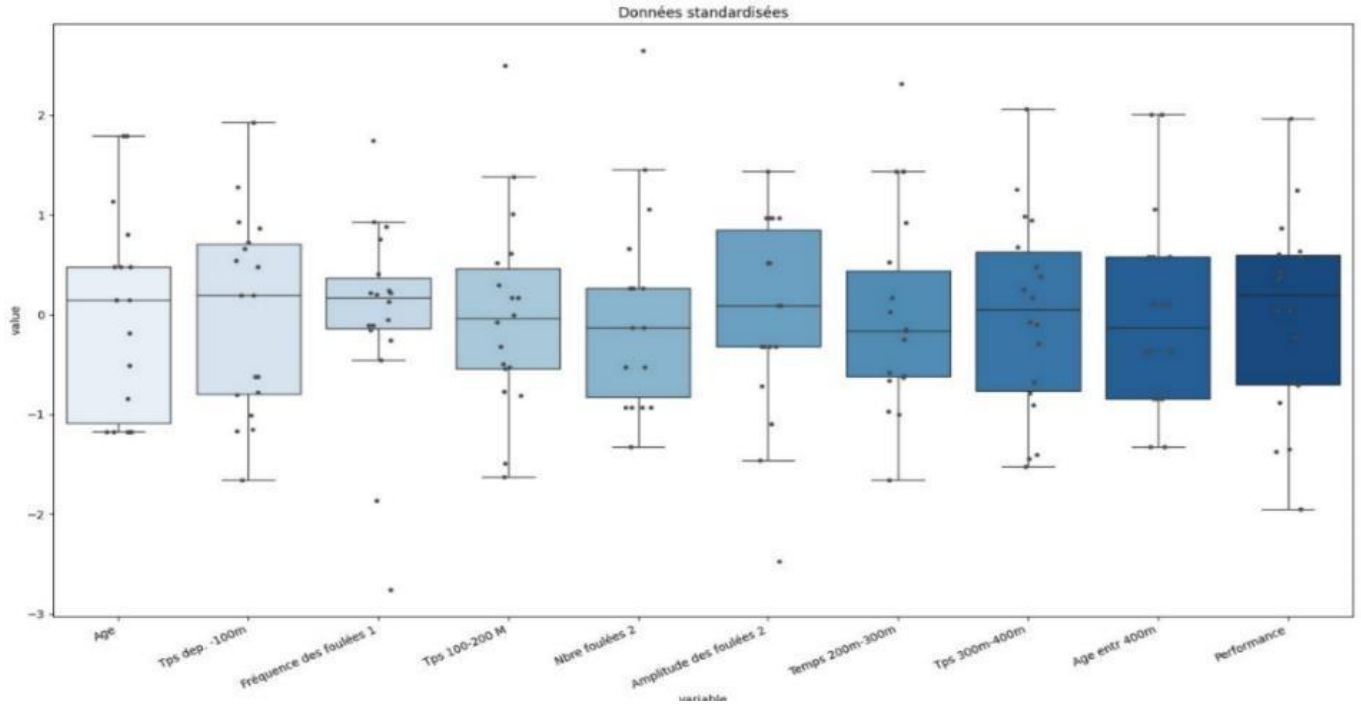

Fig. 2. Dispersion of athletes around the median on variables with significant correlation

If we take into account only the variables with a significant correlation, we end up with a better dispersion around the median, as shown by the time for the first 100m of the race.

This representation was only possible after the variables had been standardised, as the raw data contained variables with different units and sizes.

3.2 BIVARIATE ANALYSIS OF THE DIFFERENT VARIABLES USED TO MODEL THE 400M RACE

The results of our multiple linear regression analysis reveal a significant relationship between time in the first 100 metres, age and performance in the 400 metres. The coefficients obtained, ~~plotted~~ on the heatmap below, provide an indication of the relative impact of each variable on overall performance.

The coefficient (0.95) associated with the time of the first 100 m indicates the direct influence of this phase on the 400 m. Similarly, the coefficient (0.53) associated with age highlights the importance of cumulative experience in predicting performance.

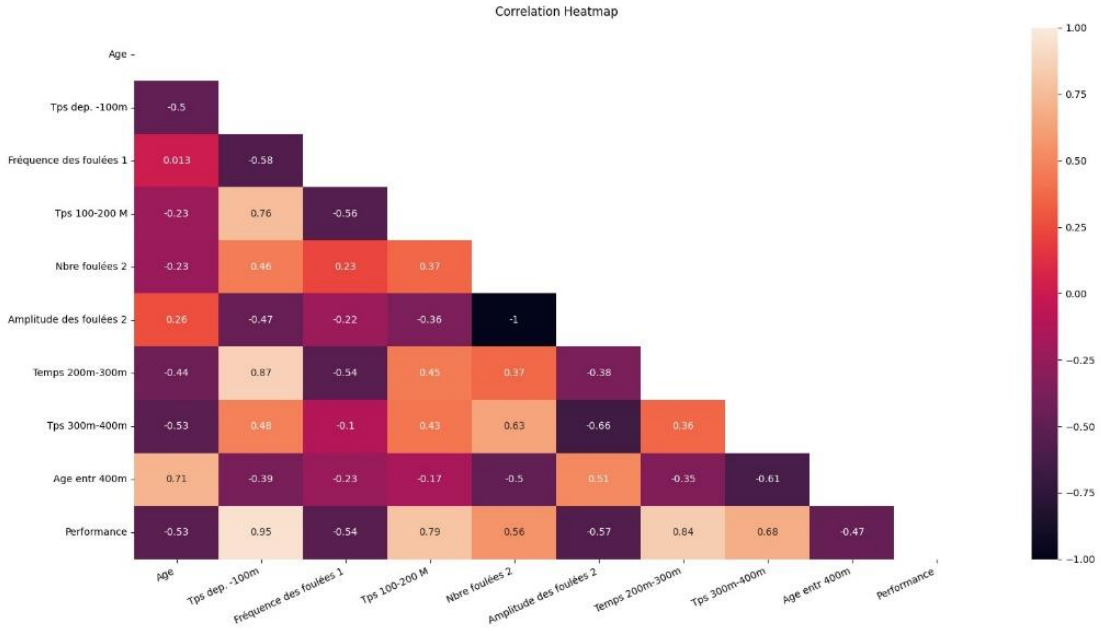

Fig. 3. Heatmap of correlations between the various variables studied and 400m time

For visual understanding, graphs relating the variables to 400 m performance will be included. These graphs will allow you to intuitively grasp the impact of the variables and visualise the general trend observed in our sample.

3.3 PERFORMANCE DEPENDS SOLELY ON THE TIME TAKEN IN THE FIRST 100 METRES

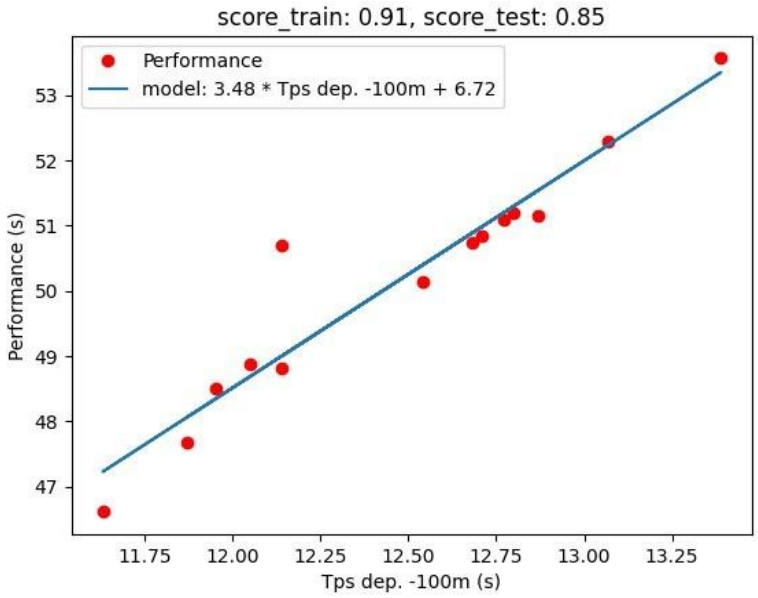

Fig. 4. 400m time prediction model based on the time achieved in the first 100m of the race

The Student's t-test indicates a p-value of 1.69 10^{-9}. The critical probability associated with the variable *time-dep.-100m* is almost zero. As this probability is less than 5%, we reject the null hypothesis H0. *time-dep.-100m* has a significant positive impact on performance.

As for the Fisher test, the critical probability is 1.17 10^{-9}, so the null hypothesis is rejected. We therefore conclude that the model is significant overall.

From our study sample (N=18), we randomly selected fourteen (14) subjects to make up the train population and the other four (4) athletes to make up the test population. The model obtained is as follows ***400m time = 3.84*time dep- 100m+6.72***

Where "*time dep-100m*" = the time taken to cover the first 100 metres of the 400m race.

Our model gave a score of 0.91 with the train population showing that the 400m results can be obtained at 91% by applying our model which gives the 400m time from the time achieved in the first 100m of the race.

We then tested our model, which gave a score of 0.85, translating into 85% reproducibility ~~of~~ the 400m time from the time of the first 100m of the race.

3.4 PERFORMANCE DEPENDS SOLELY ON THE TIME OF THE FIRST 100 METRES AND THE AGE OF THE ATHLETE

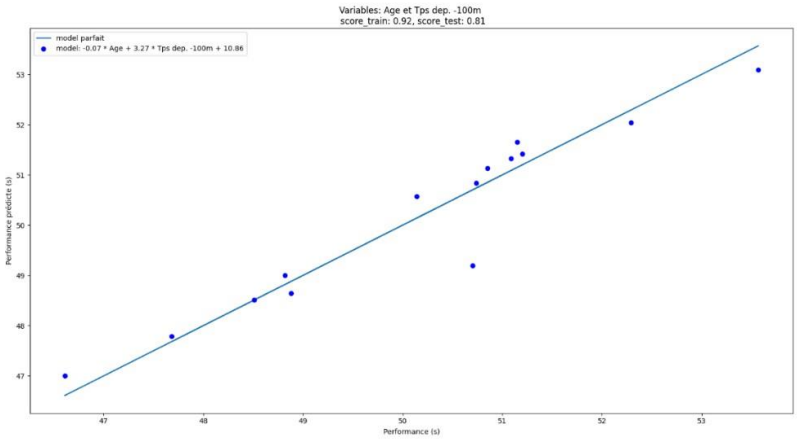

Fig. 5. Modelling of 400m performance using the first 100m of the 400m and the athlete's age

Using Student's t-test, the critical probability associated with the variable *tps-dep.-100m* is 3.4 10^{-8} and that associated with the variable age is 0.048. As these probabilities are less than 5%, the null hypothesis H0 is rejected. The *tps-dep.-100m* variable has a significant positive impact on performance, while the age variable has a significant negative impact on performance.

Using the Fisher test, the critical probability is 1.41 $10^{-8,}$ so the null hypothesis is rejected. We therefore conclude that the model is globally significant.

Following the same principle with train and test, we have made our model more complex by including the age variable. The model obtained is as follows:

400m time: (-0.07 * age) + (3.27 * tps dep. -100m) + 10.86

This gave a highly significant score ~~of~~ 0.92 for the train population, showing 92% reproducibility of the 400m performance based on the time taken between the start and the 1er 100m, compared with 0.81 for the test population.

4 DISCUSSION

Our results suggest that time in the first 100 meters and age are crucial indicators of performance in the 400 meters. Further interpretation will highlight how these factors interact to influence the success of athletes in this specific discipline.

Our results are in agreement with those of Willis R. et al (2013) [10] who show that in the 400m, the choice of initial running speed has a significant effect on the ability to maintain time during the 3ème 100m split time, which has a very strong correlation with final time (r=.90,.89, p<.01) for women and men respectively. This importance of acceleration will be echoed by Danijela Grgic (2017) [11] who explains that "*acceleration at the start is crucial for an efficient race and the quality of acceleration at the start depends on the duration and manner of the first and subsequent stages after the start*". In fact, the three athletes on the podium were already the first to complete the first 100m. Although the second 100m is very often faster and the third follows, the first remains an important element in predicting the final time.

Understanding how first 100m time and experience contribute to performance offers valuable insights for developing personalised training programmes. Coaches will be able to specifically target these aspects, thereby improving athletes' preparation for the 400m.

Although our study provides significant insights, it has certain limitations. For example, a larger sample size could be used to strengthen external validity. Future research could explore other relevant variables with a view to establishing other predictive models that would be more applicable to other athletes.

5 CONCLUSION

In summary, our modelling of the 400m using linear regression offers valuable insights into the determinants of performance in this discipline. The coefficients obtained offer tangible indications of how time in the first 100m and age contribute to success.

This study contributes to our understanding of the key factors influencing performance in the 400 metres. The practical implications for coaches and athletes provide a sound basis for improving individual performance and designing more effective training programmes.

REFERENCES

[1] R. Thomas, J. P. Eclache and J. P. Keller, «Les aptitudes motrices: Structure et évaluation,» *Journal of Fitness Research,* 01 January 1989.
[2] R. B. Alderman, Psychological Bahaviour in Sport, Philadelphia: Thompson learning, 1974.
[3] B. J. Cratty, Social dimension of physical activity, Prenticehall Inc, 1967.
[4] V. Carron, Social psychology of sport, Movement Publications, 1980.
[5] J. Weineck, Optimales training, Perimed, 1983.
[6] T. W. Calvert, E. W. Banister, M. V. Savage and T. Bach, «Model of the Effects of training on Physical Performance,» *IEEE Transaction on system man and cybernetics,* pp. 94-102, February 02, 1976.
[7] M. Franks and D. Goodman, «A systemic Approach to, analysing sport performance,» *Journal of sports sciences,* pp. 49-5, 1986.
[8] M. Arnauld, Year plans for speed and strength endurance for 400 metre runners, Athletics coach, 1989, pp. 33-44.
[9] V. Gambetta, Training and technique for the 400m dash, vol. 3, Track and field Quarterly review, 1985.
[10] R. Willis, B. Burkett and M. G. Sayers, Journal of fitness research, vol. 1, Australian Institute of fitness, 2013, pp. 40-49.
[11] D. Grgic, Dinamika trcanja u disciplini 400 metara, Zagreb: Faculty of Kinesiology, 2017.