Historical and Worldwide Trends in Ultramarathon Running: An Analysis of Participation and Performance Over Two Centuries

B218548

April 1, 2024

1 Overview

This study provides a comprehensive examination of the historical and current trends in ultramarathon participation and performance on a global scale. Analyzing a dataset [2] spanning from 1900 onward, the research focuses on identifying patterns in athlete endurance, with a spotlight on examining the influence of significant historical events such as the World Wars and the COVID-19 pandemic, alongside exploring shifts in gender participation and the trajectory of performance improvements over time. Notable findings include the resilience of the sport through adverse global events and a marked increase in female participation from the 1980s. Moreover, the research observes a performance plateau, suggesting a possible approach to human physiological limits in the sport. Focusing specifically on data from the 2010 USA ultramarathon events, the research succeeds in developing a predictive model for finishing times. This model demonstrates considerable predictive accuracy, with an R^2 score of 0.689 and a mean absolute error of 74.5 minutes. Future iterations of this research could significantly benefit from integrating detailed athlete training, nutrition, and biometric data to enhance model comprehensiveness.

2 Introduction

Context and motivation The domain of ultramarathon running offers a unique perspective for scholars examining the extremes of human performance and participation demographics. Ultramarathons, defined as any race exceeding the traditional 42.195km marathon distance, provide fertile grounds for exploring demographic trends, human exertion, and physiological limits. The intrigue of this research is motivated by a desire to delve into the complexities of ultramarathon participation and performance on a global scale, exploring how these dimensions intersect with physiological, psychological, and sociocultural determinants.

Previous work The academic field benefits from significant contributions by studies like those of Stöhr et al. (2021) [1] and Thuany et al. (2023) [3], which have meticulously examined global trends and specific country patterns in ultramarathon events. Stöhr et al. (2021) laid a foundational layer for this line of inquiry, highlighting the increasing popularity of ultramarathons in North America since the new millennium. Additionally, they noted a significant uptick in participant numbers and a steady increase in female representation. Thuany et al. (2023) conducted a thorough analysis of 100-mile ultramarathon participation and performance, distinguishing between continents and pinpointing the countries with the fastest athletes. Their study showed a dominant presence of athletes from America and Europe, with Africa producing the swiftest performances, highlighting the diverse origins of excellence in this sport.

Objectives This project aims to deepen the understanding of ultramarathon running by analyzing data from 1900 to the present, focusing on three key areas. Firstly, it examines the evolution of participation, investigating how significant historical events have influenced ultramarathon involvement and tracking changes in female participation over time. Secondly, it explores the evolution of elite performances to determine if top athletes' finishing times are approaching a plateau, suggesting potential limits to human endurance. Lastly, the project seeks to predict future performances by identifying historical trends, competition types, and participant characteristics that influence finishing times. This multifaceted analysis is expected to benefit a broad spectrum of individuals within the endurance sports community, from researchers exploring human performance boundaries to practitioners and coaches aiming to optimize training and competitive strategies.

3 Data

Data provenance The dataset for this study was sourced from Kaggle's "The big dataset of ultra-marathon running" [2], which encompasses over 7 million race records from 1798 to 2022. This expansive dataset is publicly available under the CC0: Public Domain license, ensuring its open use for academic and research purposes. Compiled from various public websites, the dataset was anonymized to protect athletes' privacy by substituting names with unique Athlete IDs, thereby maintaining data integrity while complying with data protection laws.

Data description This dataset provides a detailed account of ultra-marathon events, encompassing an extensive array of over 7 million race records (7,461,226 rows × 13 columns) from 1798 to 2022. It includes variables such as the year of the event, event dates, event name, and the distance/length of the event, which ranges from fixed distances (e.g., 50km, 100mi) to time-limited challenges (e.g., 24h, 2d). Additional details include the number of event finishers, athletes' performance times or distances, club affiliations, country of origin (denoted by codes such as USA, GBR, JPN), athletes' year of birth, gender, age categories, average speeds, and unique athlete identifiers.

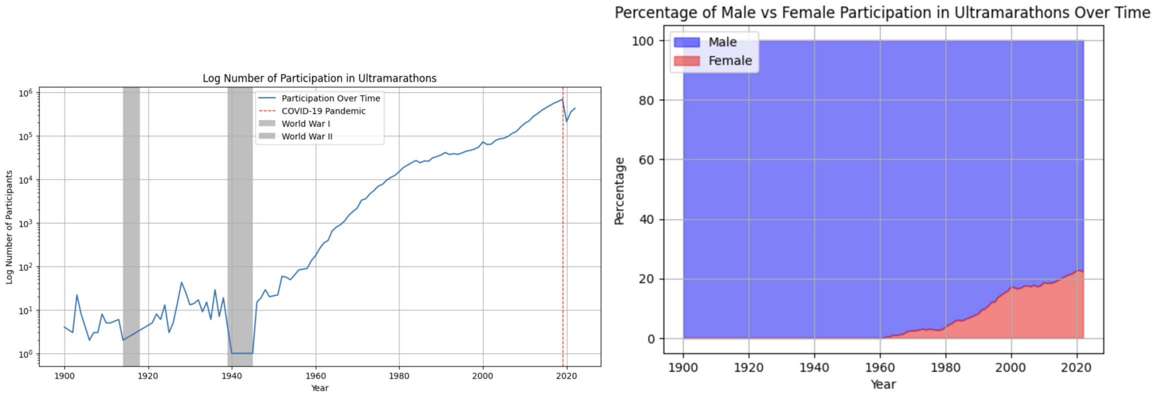
Data processing In preparing the dataset for analysis, the initial cleaning process involved retaining records from 1900 onward, focusing on modern ultra-marathon events. Records missing crucial information, such as event details, athlete performance, and demographic data, were removed to ensure data quality. The event distance/length variable underwent standardization through regular expressions to accurately convert varied race formats (kilometers or miles) into a consistent unit of measurement (kilometers) and to classify time-based events. For the specific purpose of predictive modeling, a focused subset targeting the 2010 USA Distance Events with finishers under 1200 minutes was extracted. This subset was further refined to ensure relevance by applying regex for distance standardization and excluding events not matching the defined criteria.

Feature Engineering To facilitate in-depth analyses, several new variables were engineered. The presence of club affiliation was denoted by the 'Have_club' variable, assigning a 1 for athletes affiliated with a club and 0 for those without, based on the "Athlete club" column's data availability. Athletes were also categorized by their continent of origin, translating country codes into continents and encoding this information into binary variables for each continent, such as Africa, Asia, Europe, North America, Oceania, and South America. Additionally, the "Athlete gender" variable was transformed to use numerical encoding, with 1 representing male and 0 indicating female athletes, thereby replacing the original M and F designations. These processing steps and feature engineering efforts provide a structured foundation for the subsequent exploration of ultramarathon participation and performance trends.

4 Exploration and analysis

4.1 Evolution of participation

Visualization Analysis The first graph portrays the logarithmic trend of ultramarathon participation over the last century. Notably, despite the inherent fluctuations and declines during the periods of World



(a) Logarithmic scale plot of the number of ultramarathon participants over time

(b) Stacked area chart of the percentage of male versus female ultramarathon participants over time

Figure 1: Evolution of ultramarathon participation highlighting overall growth trends and changes in gender dynamics

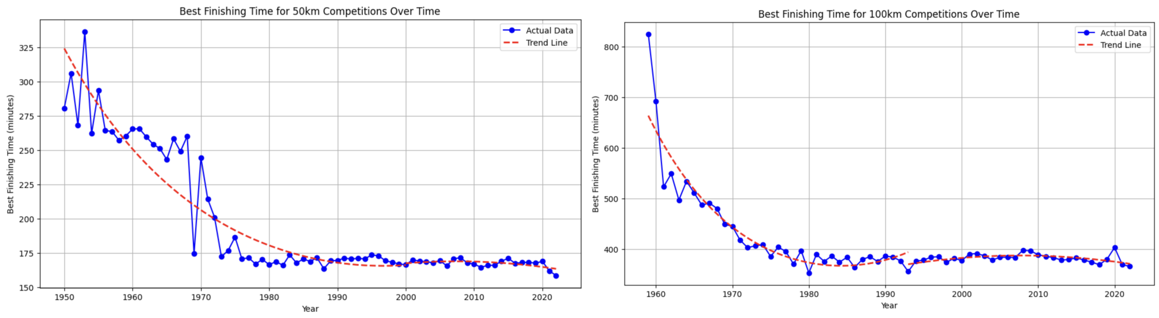
War I, World War II, and the recent COVID-19 pandemic, there has been an exponential growth in the number of participants, suggesting a strong and persistent increase from the early 1900s to the present. This logarithmic scale is crucial to understanding the trend since it accommodates a vast range of values, revealing the true exponential nature of participation growth. The second stacked area chart presents the percentage of male versus female participation in ultramarathons over time. It displays a predominantly male participation until the 1980s, after which there is a clear and steady increase in female participation. This uptick correlates with broader societal changes that have encouraged female involvement in sports historically dominated by males.

Interpretation The analysis suggests that ultramarathon running has displayed resilience and increased popularity, even in the face of global adversities. The World Wars and COVID-19 pandemic appear as temporary deterrents to participation growth but have not derailed the overall upward trajectory. The increasing female participation from the 1980s onwards is particularly noteworthy, indicating a shift towards greater inclusivity within the sport. This could be a reflection of changing social norms, the increasing visibility of female athletes, and the encouragement of female participation in the sport.

Discussion of Findings and Implications These findings are significant because they underscore the growing appeal of ultramarathons as a sport that welcomes a diverse group of athletes. They also highlight the endurance of the sport's popularity, not just as a physical challenge but as a social and cultural phenomenon that adapts and evolves through time. The surge in female participation marks a pivotal change in the sport's demographics, suggesting a future of ultramarathon running that is more diverse. Although there has been significant progress, the path to gender parity remains long. This ongoing disparity suggests more work is needed to remove the barriers to women's participation and to foster an environment that encourages a more balanced gender representation in the sport.

4.2 Evolution of elite performances

Visualization Analysis The visual data for the 50km and 100km ultramarathon events display a discernible pattern over time. The initial downward trajectory in finishing times indicates a period of rapid performance improvement. The plots show a concentration of best finishing times that steadily decrease, which the trend line reflects by descending sharply during the earlier years covered in the dataset. Post-1990, both distances show a marked leveling off in the rate of improvement, illustrated by the flattening of the trend lines, suggesting a performance plateau. Despite marginal yearly fluctuations,



(a) Best finishing times for 50km ultramarathon competitions over time

(b) Best finishing times for 100km ultramarathon competitions over time

Figure 2: The evolution of elite performance in ultramarathons

the trend lines indicate that there has not been significant improvement in the best finishing times over the last 30 years.

Interpretation The performance plateau indicates that while athletes have been getting faster, the rate of improvement has slowed down significantly. The performance times for both distances appear to stabilize, with 50km times leveling out near 175 minutes and 100km around 400 minutes. Such a plateau suggests that athletes have reached a performance saturation point under the current conditions of training, nutrition, and race strategy.

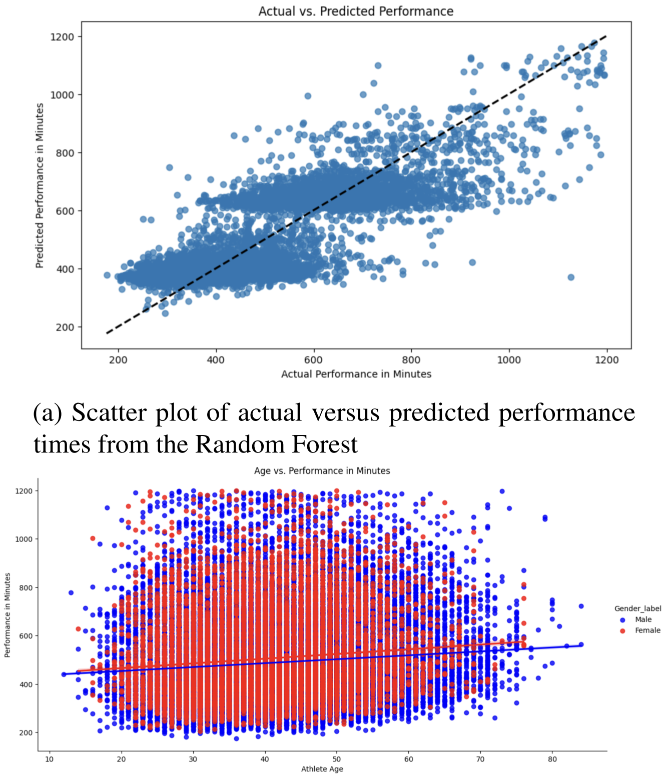
Description of Statistical and ML Methods Applied To analyze the performance data, piece-wise polynomial regression models were employed to fit the trend lines to the data, capturing the distinct phases of improvement and plateau. This approach segmented the performance data into intervals, fitting distinct polynomials to capture the dynamic changes over time so that changes in the rate of performance improvement over time can be adapted, unlike a single model that assumes a consistent pattern throughout the entire dataset. By applying this technique, the initial phase of rapid improvement in finishing times was captured by a polynomial with a steep negative slope. Conversely, to indicate a performance plateau, a polynomial with a minimal slope was fitted to represent the relatively stable finishing times.

Discussion of Findings and Implications The apparent leveling off in ultramarathon performance times has profound implications. It implies a nearing of the upper bounds of human physiological capability in long-distance running, signaling that without novel interventions or technological advances, athletes might not see substantial improvements in finishing times. The plateau also prompts a deeper examination of other avenues for performance enhancement, potentially leading to innovative breakthroughs in training, equipment, and racing tactics.

4.3 Factors to predict finishing times

The first scatter plot visualizing actual versus predicted performance, created using a Random Forest regression model, shows a dense clustering of points around the line of best fit, indicating that the model has strong predictive accuracy for a majority of the data, especially for average finishing times. Nonetheless, a notable dispersion is observed for longer finishing times, which highlights a decline in predictive precision for these instances.

This model was trained on preprocessed features including event distance, athlete age, gender, club membership, and geographical regions, and evaluated for its predictive accuracy with R^2 and MAE as the metrics. The importance analysis ranked event distance as the primary determinant of finishing times with an importance of 0.92, a logical outcome considering that longer events inherently require more time.



(a) Scatter plot of actual versus predicted performance times from the Random Forest

(b) Age versus performance scatter plot with linear regression lines

Figure 3: Analysis of ultramarathon performance prediction

Athlete age emerged as the second most significant factor, followed by gender, indicating these attributes also affect performance times notably. Club affiliation and geographical region are less influential to the model's predictions.

However, to gain more meaningful insights, the focus was shifted to the second and third most important features—athlete age and gender. A secondary scatter plot, which examines the relationship between age and performance while accounting for gender, reveals general trends: male athletes consistently show faster finishing times than female athletes, and finishing times tend to increase with the athlete's age. These trends are encapsulated in linear models that present a simplified, aggregated picture of performance variation across different demographic groups within ultramarathon participants.

The data and the model collectively underline age as a significant performance indicator after event distance, with gender also playing a noticeable role. The age-related increase in times resonates with the natural progression of human physiological capabilities, and the gender differences may point to broader physiological and potentially sociocultural dynamics in the sport.

5 Discussion and conclusions

Summary of findings The comprehensive analysis of ultramarathon data using statistical and machine learning methods has yielded several insightful findings. The study highlighted a significant increase in ultramarathon participation over the last century, with resilience shown through historical adversities such as global conflicts and the recent pandemic. A notable rise in female participation since the 1980s was observed, yet a gap still remains in reaching gender parity. Performance trends revealed a period of rapid improvement, followed by a plateau in recent decades, suggesting athletes may be approaching a physiological performance limit in ultramarathons. The Random Forest model identified event distance as the most influential factor on performance times, with athlete age and gender also being important but to a lesser extent.

Evaluation of own work The research methodology adopted a robust approach to data processing and extraction of features, leveraging advanced statistical methods and machine learning techniques. Specifically, piece-wise polynomial regression provided a nuanced understanding of the temporal evolution of ultramarathon performance times, demonstrating an acknowledgment of the sport's complex dynamics over simple linear approximations. The Random Forest model showcased commendable predictive power. It harnessed a multitude of variables, reflecting the multifaceted nature of endurance sports and the diversity of factors influencing athlete performance. This model's utility was particularly evident in its high R² value, affirming its robustness in generalizing across a broad dataset. However, its reliance on event distance as a primary predictive feature, due to its inherent and expected influence (as longer events will naturally take more time), inadvertently introduced a ceiling effect that potentially masked subtler yet impactful variables. Such dominance of a single, expected predictor hints at a modeling limitation and points to the necessity for a more discriminating feature selection strategy that could amplify the influence of less apparent but meaningful factors, like athlete ages, genders, and regions of participants.

Comparison with any other related work In comparison to related works like those of Stöhr et al. (2021) [1] and Thuany et al. (2023) [3], which also explored global trends and specific patterns in ultramarathon participation and performance, this study contributes a nuanced perspective by incorporating a broader array of factors over a longer timespan. While Stöhr et al. focused on the rise of ultramarathons in North America and Thuany et al. compared continental differences, this research expands on these findings by integrating historical events, gender shifts, and a predictive model to identify performance determinants.

Improvements and extensions Future improvements to this research would necessitate an in-depth exploration beyond the existing Kaggle dataset, seeking additional granular athlete-centric data, including training schedules, nutritional practices, and biometric indicators. The integration of such detailed variables aims to illuminate the complex influences on ultramarathon performance, offering a pathway to surmount the performance plateau that current, less-detailed datasets may not fully reveal.

Moreover, an essential extension of this work would be the rigorous statistical analysis to validate participation trends associated with significant external events, such as the World Wars and the COVID-19 pandemic. While initial observations indicated a correlation between these events and fluctuations in participation rates, a comprehensive application of statistical tests to determine the causality of these fluctuations remains unexplored. Implementing methods of hypothesis testing or causal inference could substantially strengthen the evidence that these historical occurrences significantly influence ultramarathon engagement.

Lastly, delving into the sociocultural dynamics that influence participation, especially for female athletes, stands as a critical area for future research. This exploration could inform more inclusive practices and policies within the sport. Through these advancements, the research could offer a nuanced, comparative analysis alongside existing literature, supported by an expanded dataset and a more sophisticated methodological approach. Such developments promise to make a meaningful contribution to the broader comprehension of endurance sports, highlighting paths for future investigations and practical applications in the field.

References

[1] A Stöhr et al. “An Analysis of Participation and Performance of 2067 100-km Ultra-Marathons Worldwide”. In: *International Journal of Environmental Research and Public Health* 18.2 (2021), p. 362. DOI: 10.3390/ijerph18020362.

[2] *The big dataset of ultra-marathon running*. https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running/data. A huge collection of over 7M race records registered between 1798 and 2022. Accessed: 2024.

[3] M Thuany et al. “A macro to micro analysis to understand performance in 100-mile ultra-marathons worldwide”. In: *Scientific Reports* 13.1 (2023), p. 1415. DOI: 10.1038/s41598-023-28398-2.