Clustering Aging Curves to Classify Athlete Development and Predict Career Trajectories

David Awosoga, Yushi Liu, and Samuel WK Wong

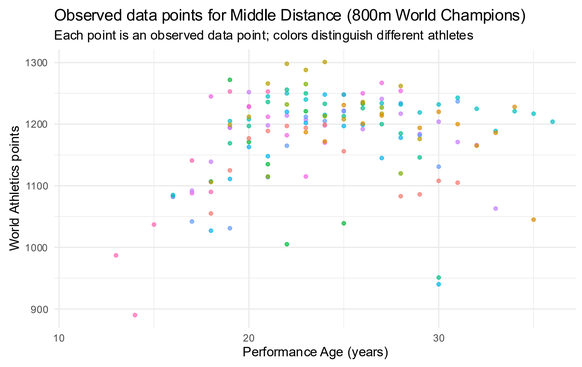
Department of Statistics & Actuarial Science, Faculty of Mathematics

Introduction

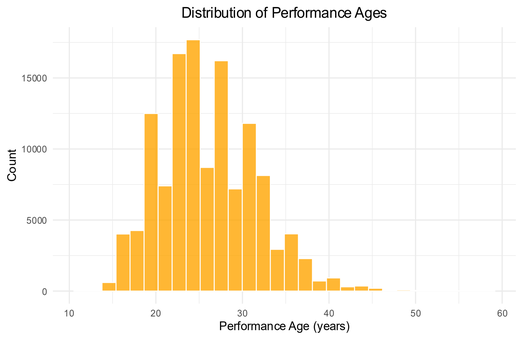
Athletes achieve peak performance at highly variable rates: some progress rapidly, while others improve gradually over longer careers. Unlike team sports, track and field results are objectively measured and comparable across events through the World Athletics scoring system[1]. By modeling aging curves using functional data analysis, we aim to classify athlete development patterns and provide predictive insights into career trajectories.

Objective

We apply Functional Principal Components Analysis (FPCA)[2] and informatively missing FPCA (imFunPCA)[3] to model athlete aging curves, correct for right tail selection bias, and extract dominant progression patterns. Athletes are clustered at both the population level and by event type using the percent change of World Athletics points to capture different rates of progression. These clusters provide predictive insights into when athletes are likely to reach elite standards based on early career performance.



Career Progression for 800m Olympic Champions, highlighting the sparse and uneven distribution of observations.



Histogram of performance ages across all athletes.

Data

We use **seasons-best performances** of Olympic track and field Athletes from World Athletics[4]. Results are standardized into **World Athletics points** using the World Athletics scoring system[4], enabling comparisons across disciplines such as sprints, jumps, and throws. Analyses cover performance ages 11 to 58. Each athlete's competition history forms an aging curve, and athletes competing in multiple events contribute separate trajectories for event-specific clustering.

Exploratory Data Analysis

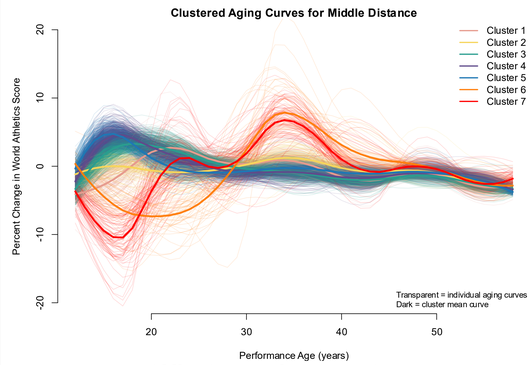
Athlete performance results show considerable irregularity. Some athletes compete frequently while others have sparse participation, leading to uneven data across ages. Career lengths also vary, with less successful athletes dropping out earlier, creating a right tail selection bias. We computed average Tokyo 2025 World Championships standards across events for men and women, allowing us to examine when athletes typically reach elite benchmarks. These patterns motivate the need for advanced modeling methods.

Methodology

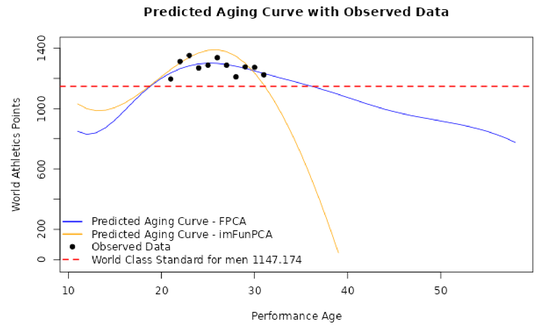
We modeled athlete aging curves using functional data analysis. FPCA was applied to smooth uneven competition results and identify dominant progression patterns. To address right tail selection bias, we used imFunPCA, which assumes athletes cannot surpass their last observed performance after exiting competition. Clustering was then performed on the percent change of World Athletics points, at both the population and event level, to group athletes with similar developmental trajectories.

Predictive Modeling Results

FPCA and imFunPCA produced smoothed aging curves that captured both dominant progression patterns and realistic late career declines. Clustering on the percent change of World Athletics points, validated through scree plots and cross validation, identified seven optimal clusters. These clusters revealed distinct developmental profiles such as early risers, steady performers, and late bloomers, providing insight into how athletes progress toward elite standards at different rates.



Middle Distance Clustering using FPCA



The predicted aging curve of Usain Bolt in the 100m.

Case Study: Usain Bolt

Usain Bolt's aging curve highlights the importance of censor aware modeling. The imFunPCA trajectory captures his sharp late career decline, while sparse FPCA smooths results but unrealistically sustains performance by ignoring censoring. This example demonstrates how imFunPCA provides a more accurate representation of athlete progression and underscores the value of accounting for selection bias in predictive modeling.

Future Extensions

Future work could extend this analysis using **Multilevel Functional Data Analysis**[5] to model multiple observations per athlete, capturing both between athlete and within athlete variability. Incorporating **training history, injury data, and external factors** may improve predictive accuracy. We are also developing an **interactive Shiny application** that visualizes FPCA based predictions of aging curves and athlete comparisons. A future goal is to integrate censor aware models such as imFunPCA, making the tool more robust for coaches and analysts.

UNIVERSITY OF WATERLOO

UWAGGS

Check out the Shiny app here!

References:

1. Awosoga, D. (2024). Peaks and primes: Do athletes get one shot at glory? *Significance, 21*(3), 6–9. https://doi.org/10.1093/jrssig/qmae038.
2. Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer.
3. Shi, H., Dong, J., Wang, L., & Cao, J. (2021). Functional principal component analysis for longitudinal data with informative dropout. *Statistics in Medicine, 40*(3), 712–724. https://doi.org/10.1002/sim.8798.
4. World Athletics. (n.d.). *World Athletics*. https://worldathletics.org
5. Crainiceanu, C. M., Goldsmith, J., Leroux, A., & Cui, E. (2024). *Functional Data Analysis with R* (1st ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781003278726.

Acknowledgements:
This work was made possible via collaboration within the University of Waterloo Analytics Group for Games and Sports (UWAGGS). A special thank you to Rithika Silva for his extensive support acquiring the data.