OPEN ACCESS

Journal of Science & Cycling
Breakthroughs in Cycling & Triathlon Sciences

Conference Abstract

Science and Cycling Conference, Lille 2025

# Clustering Professional Road Cycling Races into Distinctive Race Profiles

**Annemiek Roete [1,*], Marije Elferink-Gemser [1], Robert Lamberts [1], Matthias Kempe [2]**

**Received:** 28 Febraury 2025 **Accepted:** 12 March 2025 **Published:** 19 November 2025

[1] Department of Human Movement Sciences, University Medical Center Groningen, Groningen, The Netherlands

[2] Department of Biomechanics, Kinesiology and Computer Science in Sport, Centre for Sport Science and University Sports, University of Vienna

**Correspondence**

Annemiek Roete

Department of Human Movement Sciences, University Medical Center Groningen, Groningen, The Netherlands

a.j.roete@umcg.nl

**Keywords**

races, machine learning, physiological

## 1 Introduction

Professional road cycling races vary in format and terrain, resulting in diverse physiological demands on riders. These variations are largely attributable to the race profiles. A better understanding of race demands resulting from different race profiles can optimize cyclists' performance by tailoring training to race types. Additionally, it can support the decision making around a riders' race program, training periodization, and can enhance team composition and scouting by matching riders' strength and weaknesses to specific race profiles. Race organizers or researchers often apply their own criteria to categorize races into distinctive types. There is currently no standardized categorization for professional men's road cycling. This study aims to develop a categorization model for professional road cycling races using clustering analysis.

## 2 Material and Methods

The initial dataset comprised international races listed on the 2024 Union Cycliste International (UCI) calendar, including .WT, .Pro, .1, and .2 classifications. Features used in earlier studies with a physiological focus which categorized races into distinct types were summarized and used as features in the clustering analysis. Relevant data to quantify those features were obtained from

ProcyclingStats, an open-access database. The features identified in earlier research included race distance, vertical meters, number of climbs, total uphill distance, vertical meters of all climbs, vertical meters of climbs positioned in the last 10km and, vertical meters located at the finish. To enhance the accuracy and stability of the clustering process, noise reduction was performed by identifying races that had been explicitly categorized by organizers, as these were more likely to be well-defined and distinctive. Additionally, all data was thoroughly checked. Incomplete and or incorrect data was removed. The clustering tendency of the dataset was assessed using the Hopkins statistics. Three clustering algorithms were then applied: KMeans, hierarchical clustering with genie linkage and noise detection, and the Ordering Points To Identify Cluster Structure (OPTICS) algorithm. The Silhouette score was used to find the best model per algorithm. To compare those models, a subset of labeled races provided by various race organizers were used as ground truth.

## 3 Results

In total, 1,020 race days were reviewed for inclusion into the clustering dataset, with 276 race days being included. Hopkins statistics was 0.70, indicating the dataset is clusterable. KMeans clustering produced five distinct clusters: time trial, flat, hilly, semi-mountain and mountain. Hierarchical clustering also identified five clusters corresponding to time trial, flat, hilly, semi-mountain, mountain. OPTICS generated 3 clusters, corresponding mostly to time trial and 2 flat clusters (with varying vertical meters). 143 races were used as ground truth (consisting out of flat, hilly, semi-mountain, mountain and time trials). KMeans outperformed hierarchical clustering and OPTICS across all ground truth labels.

## 4 Discussion

The differences in clustering results reflect the inherent variations in the algorithms' methodologies. KMeans demonstrated the best performance compared to hierarchical clustering and OPTICS. While hierarchical clustering and OPTICS are better suited for handling complex data structures, their results suggest a tendency to overfit the dataset in this context, especially given the inherent variability in road cycling races, even within the same category. The findings of this study suggest, that based on clustering, race profiles of men professional cycling races may be divided into flat, hilly, semi-mountain, mountain, and time trial race profiles. Using such a unified categorization of race profiles allows for a comparison of different races and studies, providing a comprehensive overview of all professional cycling races. Additionally, this standardized classification enables a deeper analysis of race demands by recognizing variations in race profiles.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.