



## Velocity zone classification in elite women's football: where do we draw the lines?

Laurence A. F. Park, Dawn Scott & Ric Lovell

**To cite this article:** Laurence A. F. Park, Dawn Scott & Ric Lovell (2019) Velocity zone classification in elite women's football: where do we draw the lines?, *Science and Medicine in Football*, 3:1, 21-28, DOI: [10.1080/24733938.2018.1517947](https://doi.org/10.1080/24733938.2018.1517947)

**To link to this article:** <https://doi.org/10.1080/24733938.2018.1517947>



Published online: 12 Sep 2018.



Submit your article to this journal [↗](#)



Article views: 1336



View related articles [↗](#)



View Crossmark data [↗](#)





Citing articles: 31 View citing articles [↗](#)

ARTICLE



## Velocity zone classification in elite women's football: where do we draw the lines?

Laurence A. F. Park <sup>a</sup>, Dawn Scott<sup>b</sup> and Ric Lovell <sup>c</sup>

<sup>a</sup>School of Computing, Engineering and Mathematics, Western Sydney University, Sydney, Australia; <sup>b</sup>High Performance Department, US Soccer Federation, Chicago, IL, USA; <sup>c</sup>School of Science and Health, Western Sydney University, Sydney, Australia

### ABSTRACT

**Objectives:** This study aims to develop generic velocity thresholds for the analysis of external load data collected in international women's football matches.

**Methods:** Doppler-derived recordings of instantaneous velocity and acceleration were collected (10 Hz GPS) from 27 international female football players during 52 international matches between 2012 and 2015. Data were examined with *k*-means, Gaussian mixture model (GMM), and Spectral Clustering methods to identify four velocity zones, in each completed half of match-play (277 observations). Spectral Clustering was also performed with 4 different smoothing parameters ( $\beta$  values of 0, 0.001, 0.01, and 0.1). Linear-mixed modelling was used to determine generic squad thresholds, accounting for the within-subject variation.

**Results:** *k*-means and GMM generated low transition velocities, which had limited logical validity and deemed not fit for purpose. Spectral Clustering with a  $\beta$  value of 0.1 derived thresholds that differed from the various methods adopted in existing literature and industry practice, yet providing a rigorous, acceptable, and feasible determination of velocity thresholds.

**Conclusion:** Velocities of 3.46 (12.5 km h<sup>-1</sup>), 5.29 (19.0 km h<sup>-1</sup>), and 6.26 m s<sup>-1</sup> (22.5 km h<sup>-1</sup>) are recommended as entry criteria into high, very-high velocity, and sprinting locomotor categories, respectively, for the purpose of external load assessments in elite women's football.

### ARTICLE HISTORY

Accepted 26 August 2018

### KEYWORDS

GPS; football; velocity thresholds; data-mining; female

## Introduction

External load monitoring in football training and competition has become ubiquitous at professional levels around the globe (Akenhead and Nassis 2016). It is commonly used to track both men's and women's loads (Di Salvo et al. 2007; Rampinini et al. 2007; Bradley et al. 2014a; Datson et al. 2017; Trewin et al. 2018), but also in other populations such as elite-youth players (Buchheit et al. 2010; Harley et al. 2010) and even amateur cohorts (Dellal et al. 2011). In an attempt to capture meaning from the highly stochastic nature of football activity, the distances covered in velocity and acceleration are routinely binned in categories, or zones. The application of generic criteria to denote these zones facilitates benchmarking the work performed between different players, positional roles, and competition standards. Researchers have also used this technique to explore the impact of contextual factors such as quality of the opposition (Rampinini et al. 2007), environmental conditions (Nassis et al. 2015), and competition standard (Bradley et al. 2010) upon physical match performances. While the use of "zones" or "bins" in time-motion analysis is universal, the junctures representing the transition between locomotor or intensity classifications are rarely justified.

Early time-motion analyses in football developed locomotor classifications (e.g., standing, walking, jogging, running, sprinting) defined by movement characteristics evaluated by single observers. Mean squad velocities (Bangsbo et al. 1991; Mohr et al. 2003) were assigned to approximate both total distance covered, and the distribution of work performed in each movement category.

The assigned velocities were derived from video recordings of nine players performing specific activities ranging from walking to sprinting, with the mean velocities determined for each movement category (Bangsbo et al. 1991). With the evolution of automated player tracking systems, which enabled instantaneous recordings of velocity at higher sampling resolutions (10–25 Hz), the velocity zones adopted in both industry practice and academic research (Di Salvo et al. 2007; Rampinini et al. 2007) seemed to reflect those previously used in the early locomotor category approaches to time-motion analysis (Bangsbo et al. 1991); albeit specific justification for the zone thresholds was not provided.

The theoretical value of categorizing velocity zones using standardized and arbitrary zones between different players is questioned, considering the individual nature of both the exercise-intensity continuum (Lovell and Abt 2013; Hunter et al. 2015; Scott and Lovell 2018), and self-selected transitions between locomotor categories (Siegle and Lames 2010). These issues become particularly apparent when applying velocity transitions, taken from elite-male players, to physical match performance data taken from lower participation standards, female cohorts, and youth players. The typical approach has been to lower the thresholds for any given locomotor category to reflect the typically lower fitness capacities (Mujika et al. 2009; Harley et al. 2010) and physical match performances in these populations (Buchheit et al. 2010; Bradley et al. 2014a). In the case of elite-female football, some discourse exists regarding the velocity thresholds to adopt (Bradley et al. 2014a; Bradley and Vescovi 2015; Datson

et al. 2017), thus further investigation is warranted (Sweeting et al. 2017b).

Given the challenges of assigning velocity zone criteria for different populations, based on either research precedent (Mohr et al. 2003) or fitness capabilities (Harley et al. 2010; Bradley and Vescovi 2015; Hunter et al. 2015), an alternative approach is to examine the distribution of the velocity data in retrospective fashion according to data-mining techniques (Sweeting et al. 2017a, 2017b). This approach was taken by Dwyer and Gabbett (Dwyer and Gabbett 2012), who fitted four Gaussian curves to the instantaneous velocity data from a range of male and female team-sports matches. The intersections between adjacent Gaussian curves were used to demarcate velocity zones. While pioneering work, the velocity zones recommended from this study have largely not been applied in team-sports research, and are questioned on the following grounds: 1) data were collected using 1 Hz GPS, which, with recent enhancements in sampling frequency, has been found to be inaccurate (Jennings et al. 2010) and underestimates sprinting distance (Randers et al. 2010); 2) the female football population sampled was not elite, reflected by the modest running velocity assigned to classify sprinting ( $5.4 \text{ m s}^{-1}$ ); 3) the low sample size ( $n = 5$ ; 25 female football match observations) may not reflect the variation in physical match performances between matches (Gregson et al. 2010) and positional roles (Datson et al. 2017); and 4) there is no evidence to suggest that the velocities within each zone follow a Gaussian distribution. More recently, Sweeting et al. (Sweeting et al. 2017a) adopted *k*-means clustering to determine movement sequences in netball, deriving velocity bins from cluster centroids. However, both *k*-means clustering and the Gaussian mixture model (GMM) assume that the instantaneous velocity data are independent and uncorrelated in nature (i.e., not related between successive data entries). We know that the instances of each velocity sequence are not independent (providing the continuity of the sequence). These two issues suggest that there is limited foundation to using these data-mining techniques, other than that they can provide a set of zones.

Accordingly, the aim of the current study was to take a data-mining approach starting from first principles, to develop velocity thresholds for elite-female football players. We also derived thresholds via other data-mining techniques used in previous work (*k*-means clustering and Gaussian mixture modeling), in an attempt to evaluate their utility. Considering the growth and increasing professionalization of women's football, combined with the proliferation of external load monitoring, data of this nature are necessary to inform both academic research and applied practice.

## Materials and methods

### Subjects

Physical match data taken from a squad of 27 international women's football players (Age:  $24.6 \pm 3.8$  years; Stature:  $168.9 \pm 4.8$  cm; Body Mass:  $63.0 \pm 4.2$  kg; YoYo IR1:  $1760 \pm 240$  m; Maximal Oxygen Uptake:  $54.9 \pm 3.2$  ml kg  $\text{min}^{-1}$ ; Peak Speed:  $8.06 \pm 0.42$  m  $\text{s}^{-1}$ ), who participated in 52 different matches between 2012 and 2015 was used for this analysis. The squad was ranked #1 in the FIFA World Ranking

over the assessment period and won the FIFA Women's World Cup in 2015. The matches were noncompetitive, scheduled in preparation for major tournaments. The procedures were approved by an institutional Human Research Ethics Committee.

### Methodology

Outfield players wore 10 Hz GPS devices (MinimaxX S4, Catapult Sports, Australia) between the scapulae in a neoprene undergarment. At the time of data collection, 10 Hz GPS were considered the most reliable and valid devices available for both linear and team-sports running (Scott et al. 2016). A minimum of 45 min of playing time was required for each eligible match observation to discount the high work-rates associated with substitute players (Bradley et al. 2014b). Accordingly, analysis was performed on each playing half (i.e., a player playing a full 90+ min game would provide two match observations), inclusive of added time, to yield the maximum data input to the model. Doppler-derived instantaneous velocity data was processed using the manufacturer's "Intelligent Motion Filter" and the minimum effort duration was set at 0.2 secs. Thereafter, the data velocity data was exported from the manufacturer's software (Sprint, version 5.1.7, Catapult Sports, Australia) to R (version: 3.3.3, R Foundation for Statistical Computing, Vienna, Austria) for further processing. According to proposed GPS reporting standards (Malone et al. 2017), preliminary data filtering excluded instantaneous data where the number of connected satellites was less than 8 (range: 8–14 satellites), or the horizontal dilution of precision was greater than 2.0. Data was also treated as missing values where acceleration ( $>6 \text{ m.s}^{-2}$ ), deceleration ( $<-6 \text{ m.s}^{-2}$ ) and velocity ( $>10 \text{ m.s}^{-1}$ ) exceeded reasonable capabilities for the sample, based upon kinematic data taken from 100-m sprint athletes (di Prampero et al. 2005). Match observations consisting of erroneous data that accounted for more than 3% of the total playing time were discarded from further analysis (31 match halves); a criterion applied in consideration of elite female peak high speed running demands ( $45 \text{ m.min}^{-1}$ ; Datson et al. 2017), and the match file durations ( $\sim 45$  mins). The total number of match observations analyzed was 277, with a median of 8 per player (mean: 10.3; range: 1–29).

### Defining velocity zones

Before partitioning the velocity data into zones, we defined zone parameters in alignment with the relevant literature (Rampinini et al. 2007; Hunter et al. 2015; Datson et al. 2017). Since the distance covered at higher velocities are the most commonly used physical performance indicators (Akenhead and Nassis 2016), we proposed a classification system that amalgamated standing, walking, and slow jogging locomotor activities in a low velocity zone (LVR), and further sub-categorized higher running distances into high (HVR), very-high velocity running (VHVR) and sprinting (SPR). The four zones are created by identifying three velocity transition thresholds  $v_l$ ,  $v_h$  and  $v_v$ , as shown in Table 1.

The early pioneering work in football time-motion analysis qualified zones according to locomotor activities such as

**Table 1.** Criteria for velocity zone categorization.

Category	Lower Bound	Upper Bound
LVR	0	$v_l$
HVR	$v_l$	$v_h$
VHVR	$v_h$	$v_v$
SPR	$v_v$	$10\text{ms}^{-1}$

LVR = low-velocity running; HVR = high-velocity running; VHVR = very-high velocity running; SPR = sprinting

walking, jogging, running and sprinting (Bangsbo et al. 1991; Mohr et al. 2003), and quantified distances covered based on the players approximate movement velocity. Accordingly, we defined a zone as a velocity interval that a player has a natural tendency to use for any given locomotor category, while the zone boundaries should be located between these velocity regions. Players may change zones and hence move at a velocity on a zone boundary, but over time, we should find that a player's velocity will transition within each zone for the majority of an activity bout. Using this notion of a zone, we define the zone boundaries as the set of velocities that are traversed through the least. In the next section, we formalize this using mathematical notation.

### Determining zones from sequence data

Given a velocity sequence  $\mathcal{V}$  of length  $N$  containing the ordered velocities  $v_1, v_2, \dots, v_N$ , where  $v_i$  is the velocity at the  $i$ th time point, we located the boundaries  $v_l, v_h$  and  $v_v$ , such that the velocity traversals were minimal.

For example, to locate only two zones,  $A$  and  $B$ , they are found as the subsets that minimize the sum of traversals:

$$\sum_{v_a \in A, v_b \in B} t(v_a, v_b)$$

where the sum is over each combination of velocities  $v_a$  and  $v_b$  from the zones  $A$  and  $B$ , and  $t(v_a, v_b)$  is the number of transitions from  $v_a$  to  $v_b$  and  $v_b$  to  $v_a$  in  $\mathcal{V}$ . This is known as the minimum cut problem, aiming to identify the best partitioning of a graph that minimizes the sum of the weight of the edges between the partitions. This problem can be solved using Spectral Clustering. Spectral Clustering has been successfully used for data segmentation in other fields of research (Park et al. 2009, 2016) but to the best of our knowledge, has not been used in sport. The Spectral Clustering algorithm generalizes to provide any number of partitions, and was adopted here to compute four partitions in the velocity transition data, providing the three velocity zone boundaries.

### Zones from velocity data

The Spectral Clustering algorithm is designed for discrete or categorical data, thus velocity data was prepared by quantizing the velocity values into uniform width bins. The velocity bins in this analysis were of width  $0.1 \text{ m}\cdot\text{s}^{-1}$  ranging from  $0 \text{ m}\cdot\text{s}^{-1}$  to  $10 \text{ m}\cdot\text{s}^{-1}$ , providing 100 velocity bins. The transitions between each velocity bin were recorded from the match velocity data and provided to the Spectral Clustering algorithm.

### Transition smoothing

Spectral Clustering uses the velocity bin transition data, treating the velocity bins as categories, and disregards the ordering of the velocity bins. Accordingly, clustering may result in interleaved velocity zones (i.e., Spectral Clustering might classify LVR as 0 to 2  $\text{m}\cdot\text{s}^{-1}$  and 3 to 4  $\text{m}\cdot\text{s}^{-1}$  with the gap being classified as one of the other zones). To enforce that the Spectral Clustering delivered ordered partitions (i.e., the first partition contains velocity bins that are less than the second partition, and the second are less than the third), neighboring velocity bin transitions (between bin  $n$  and  $n + 1$ ) were artificially increased using a smoothing factor  $\beta$  (with range 0 to 1, where  $\beta = 0$  provides no smoothing, and setting  $\beta = 1$  ignores the raw data and only smoothing information is used). This style of smoothing is commonly used to ensure ergodicity (Park and Kotagiri 2011; Park and Simoff 2013). For the purposes of identifying appropriate velocity zones for women's football, the smoothing parameter should be set to a value close to zero. In this study,  $\beta$  values of 0, 0.001, 0.01, and 0.1 were applied.

### Comparison to other data-mining techniques

The  $k$ -means and GMM thresholds (LVR, HVR, VHVR, SPR) were computed using the sample of player's velocities. Note that these two methods ignore the sequence of velocities, treating each point as being independent of each other. Each set of velocities has a cluster spike at zero. Therefore, the zero values were removed before computing the  $k$ -means and GMM thresholds.

### Analysis

The velocity zones were computed according to the different data-mining techniques. The distances covered in each zone were calculated from instantaneous velocity and acceleration data using the following equation of motion:

$$\text{Instantaneous Distance (m)} = v_i t + \frac{1}{2} a_i t^2 = \frac{v_i}{10} + \frac{a_i}{200}$$

assuming a 10 Hz signal ( $t = 0.1$ ), where,  $v_i$  is the Doppler velocity, and  $a_i$  is the Doppler-derived instantaneous acceleration.

For comparison, the distances covered in generic velocity zones as adopted in previous research in women's football were also determined (see Table 2). We applied the thresholds typically used in elite-men's football and previously transposed to examine international women's activity profiles

**Table 2.** Generic thresholds used in existing women's football research.

Velocity Zone	Generic <sub>male</sub> ( $\text{m}\cdot\text{s}^{-1}$ )	Generic <sub>gy</sub> ( $\text{m}\cdot\text{s}^{-1}$ )
LVR	<4.0	<3.33
HVR	$\geq 4.0 - 5.49$	$\geq 3.34 - 4.44$
VHVR	$\geq 5.50 - 6.99$	$\geq 4.45 - 5.55$
SPR	$\geq 7.0$	$\geq 5.56$

LVR = low-velocity running; HVR = high-velocity running; VHVR = very-high velocity running; SPR = sprinting.

(Generic<sub>male</sub>) (Datson et al. 2017), together with zones analogous to those recommended by Bradley and Vescovi (Bradley and Vescovi 2015) (Generic<sub>BV</sub>).

Velocity thresholds and the distances covered in each derived zone were log-transformed to reduce nonuniformity error. Differences in the distances covered in each zone according to the different data-mining approaches were examined using linear-mixed models (IBM SPSS version 23.0, Armonk, NY), using random effects to model the within-subject variation considering the different match observations recorded for each player. P-values generated from least-squared difference *post hoc* tests, in combination with back-transformed estimated marginal mean effect statistics, were imputed into a spreadsheet (Hopkins 2007) to derive magnitude-based inferences. The magnitude of the effect was classified as small, moderate, large, very-large, or extremely large according to standardized fractions (0.2, 0.6, 1.2, 2.0, and 4.0) of the between-subject standard deviation, calculated from the distances calculated according to the Generic<sub>male</sub> zones. Inferences were determined from the disposition of the 90% confidence interval for the mean difference in reference to the standardized thresholds (*likely* = > 75%; *very-likely* = > 95%; *most-likely* > 99.5%), but regarded as unclear if the confidence intervals overlapped both positive and negative thresholds by 5% (Batterham & Hopkins, 2006). Data are reported as the back-transformed estimated marginal means with corresponding 90% confidence intervals.

## Results

The set of velocity thresholds computed using Spectral Clustering, *k*-means, a GMM (as used in Dwyer and Gabbett 2012), Generic<sub>male</sub> and Generic<sub>BV</sub> are shown in Table 3. The table also provides the proportion of sequences that were removed from the analysis, due to providing an invalid partitioning of the sequence. With no smoothing ( $\beta = 0$ ), 28.6% of the transition velocities generated by the Spectral Clustering were not useable, whereas only 3.2% of the data was removed when adopting  $\beta = 0.1$ , and the velocity at zone transitions were similar.

The distribution of the three Spectral Clustering velocity thresholds per player is shown in Figure 1. The plots show that similar players appear at the lower and upper end of the median ordering.

The distances covered in each zone according to the different methods are provided in Table 4. As expected, the lowered velocity thresholds with Generic<sub>BV</sub>-derived greater HSR (*likely* moderate), VHHR (*likely* large) and SPR distances (*very-likely* very-large) covered versus Generic<sub>male</sub>. *k*-means and GMM re-distributed the distances covered in each zone and with *very-likely* large to most-likely extremely large differences versus the other techniques. Spectral Clustering increased the proportion of HSR distance covered (*likely* small – *very-likely* moderate), but generated a higher- and lower-SPR versus Generic<sub>male</sub> and Generic<sub>BV</sub>, respectively. The different smoothing approaches for Spectral Clustering did not alter the distribution of distances covered in each zone.

## Discussion

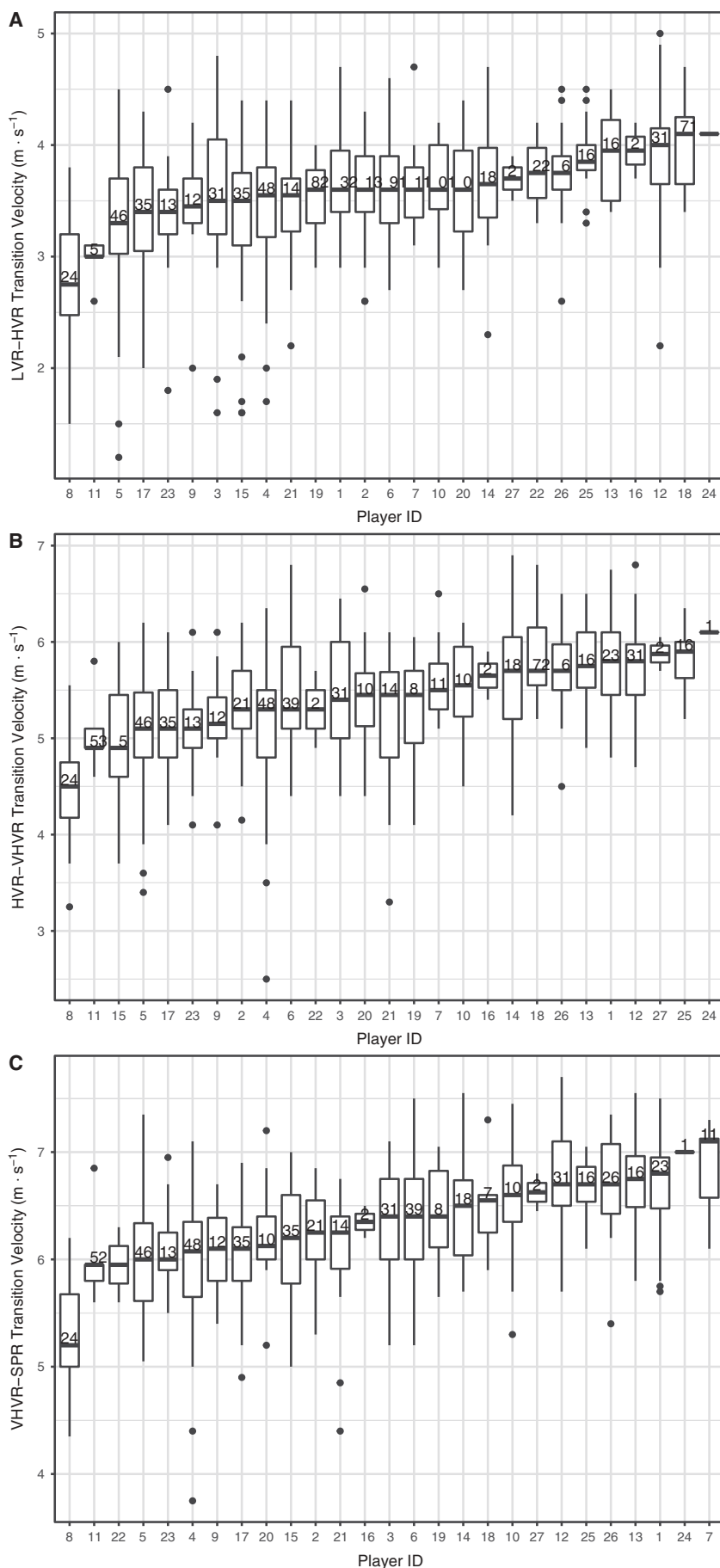
The aim of this study was to take a data-mining approach to establishing new appropriate velocity thresholds for elite women's football. We also compared the distances covered in different velocity zones according to approaches already used in the literature, together with the new zones generated retrospectively from a range of data-mining techniques. The key findings from this study were: A) Gaussian mixture modeling and *k*-means clustering generated comparatively low thresholds, deemed not fit for practice; and B) Spectral Clustering derived new squad-mean zones that were subtly different to those adopted in the research literature, yet generated meaningful differences in the distances covered in HVR, VHVR and SPR zones.

Generic velocity thresholds are universally administered in external load monitoring in sporting contexts (Akenhead and Nassis 2016). This approach permits the evaluation of training and competition work-loads, and comparisons between individuals, exposures, and squads as required by the user. Historically, the scientific justification underpinning generic zone criteria has been absent, and seemingly based on a single pilot study (Bangsbo et al. 1991) conducted on a small squad of elite-male football players prior to the evolution in athlete tracking technologies. Accordingly, as external load tracking has proliferated to other participation standards, discussion regarding the application of population- (Harley et al. 2010; Bradley and Vescovi 2015; Datson et al. 2017) or player-specific (Lovell and Abt 2013; Hunter et al. 2015) criteria to

**Table 3.** Velocity thresholds derived from the various data-mining approaches (Spectral Clustering, *k*-means, GMM), as compared to those used in previous research.

Method	Removed (%)	HVR (m·s <sup>-1</sup> )	VHVR (m·s <sup>-1</sup> )	SPR (m·s <sup>-1</sup> )
Generic <sub>male</sub>	NA	4	5.5	7.0
Generic <sub>BV</sub>	NA	3.34	4.45	5.56
<i>k</i> -means	0	1.05 (1.03–1.07)	2.10 (2.07–2.12)	3.60 (3.56–3.63)
GMM	0	0.56 (0.55–0.57)	1.53 (1.51–1.54)	3.05 (3.02–3.08)
Spec Clust ( $\beta = 0.1$ )	3.2	3.46 (3.40–3.52)	5.29 (5.23–5.35)	6.26 (6.20–6.32)
Spec Clust ( $\beta = 0.01$ )	16.6	3.54 (3.47–3.60)	5.38 (5.32–5.45)	6.30 (6.24–6.37)
Spec Clust ( $\beta = 0.001$ )	17.5	3.56 (3.49–3.63)	5.39 (5.32–5.45)	6.30 (6.23–6.37)
Spec Clust ( $\beta = 0$ )	28.6	3.58 (3.51–3.65)	5.41 (5.34–5.48)	6.27 (6.20–6.33)

LVR = low-velocity running; HVR = high-velocity running; VHVR = very-high velocity running



**Figure 1.** The distribution of the three Spectral Clustering velocity thresholds for transitions determined between LVR-HVR (a), HVR-VHVR (b), and VHVR-SPR (c) in each individual player. The boxes represent the IQR (25th to 75th percentile) of the players' transition velocities, and the whiskers represent the range. The number in each box denotes the number of player match observations. Outliers, defined as greater than 1.5 IQR from the 1st or 3rd quartile, are represented by closed circles.

**Table 4.** Distances covered in each velocity zone (m) according to the different approaches used to retrospectively generate thresholds and those commonly applied in research literature.

Method	LVR	HVR	VHVR	SPR
Generic <sub>male</sub>	3587 <sup>a</sup> (3516–3659)	434 <sup>a</sup> (419–449)	100 <sup>d</sup> (95–104)	19 <sup>a</sup> (18–20)
Generic <sub>BV</sub>	3162 <sup>b</sup> (3100–3226)	589 <sup>a</sup> (568–610)	254 <sup>a</sup> (244–266)	101 <sup>a</sup> (96–107)
<i>k</i> -means	559 <sup>a</sup> (548–570)	1312 <sup>a</sup> (1124–1207)	1402 <sup>a</sup> (1343–1463)	784 <sup>a</sup> (743–828)
GMM	135 <sup>a</sup> (132–138)	1166 <sup>a</sup> (1124–1207)	1591 <sup>a</sup> (1524–1661)	1147 <sup>a</sup> (1087–1210)
Spec Clust ( $\beta = 0.1$ )	3265 (3200–3331)	699 <sup>c</sup> (675–724)	106 <sup>d</sup> (101–110)	36 <sup>c</sup> (34–38)
Spec Clust ( $\beta = 0.01$ )	3308 (3242–3374)	670 <sup>c</sup> (647–694)	94 <sup>d</sup> (90–98)	34 <sup>c</sup> (32–36)
Spec Clust ( $\beta = 0.001$ )	3328 (3262–3395)	657 <sup>c</sup> (634–681)	92 <sup>d</sup> (89–97)	34 <sup>c</sup> (32–26)
Spec Clust ( $\beta = 0$ )	3334 (3268–3402)	655 <sup>c</sup> (632–679)	87 <sup>d</sup> (84–91)	37 <sup>c</sup> (35–39)

<sup>a</sup> denotes difference versus all other methods; <sup>b</sup> denotes difference versus Spectral Clustering ( $\beta = 0.01, 0.001, 0$ ) and Generic<sub>male</sub>; <sup>c</sup> difference versus Generic<sub>male</sub>, Generic<sub>BV</sub>, *k*-means, and GMM; <sup>d</sup> difference versus Generic<sub>BV</sub>, *k*-means, and GMM. LVR = low-velocity running; HVR = high-velocity running; VHVR = very-high velocity running; SPR = sprinting

facilitate interpretation of training and competition has emerged. Where studies have adjusted generic zone thresholds for female or youth squads, the common approach or recommendation has been to use cohort-specific physical characteristics to inform velocity transition criteria (Harley et al. 2010; Bradley and Vescovi 2015). For example, Bradley and Vescovi (Bradley and Vescovi 2015) recommended velocity thresholds based upon fitness data reported in the literature, in combination with data collected from a women's collegiate football squad. However, these recommendations have not been administered in research or applied practice on the basis that the data which derived the thresholds failed to reflect the athletic characteristics of elite-standard players (Datson et al. 2017). While using squad-mean velocities corresponding to physical and/or physiological characteristics to inform velocity zones has logical validity, in practice this technique is limited as squad zones may differ both within and between seasons, and are subject to variation when players transition between squads or are unavailable for physical screening.

A feasible alternative for determining population-specific velocity zones is to retrospectively examine latent properties in their external load data using data-mining techniques (Sweeting et al. 2017b). Given the availability of extensive player tracking data-sets which are collected routinely during professional competition, data-mining presents a feasible, affordable, and theoretically robust approach to deriving population-specific velocity zones. To the best of our knowledge, Dwyer and Gabbett (Dwyer and Gabbett 2012) were the first to adopt this approach to external load data, using GMMs to fashion sport- and gender-specific velocity thresholds for a range of team-sports. This approach assumes that the data points are generated from a user-defined number of Gaussian distributions with unknown parameters. The estimated marginal mean thresholds identified via GMM in our study differed considerably from Dwyer and Gabbett (Dwyer and Gabbett 2012). Principally, the “sprint” threshold derived was lower (3.05 vs 5.4 m·s<sup>-1</sup>), which may reflect between study differences in the GPS sampling frequency or the number of female

football match observations. Notwithstanding, the GMM does not consider the sequential pattern of consecutive velocity data-points, and as discovered in the current study, the limited observations of velocities typically associated with sprinting incorrectly assumes a Gaussian distribution. Accordingly, the underlying assumptions of the GMM preclude their utility in determining appropriate velocity zones in football, and our findings support this contention.

Application of *k*-means clustering for categorizing velocity data has also been suggested (Sweeting et al. 2017b), and used to identify common movement sequences in netball (Sweeting et al. 2017a). The *k*-means algorithm operates iteratively to assign each observation to one of *k*-specified clusters (four in this study) based on the closest centroid. This data-mining technique is also limited by the discreet manner in which each velocity observation is treated; however, it has no underlying assumptions regarding the distribution of the data. The velocity zones derived were low compared to existing methods, with transition velocities approximately 0.5 m·s<sup>-1</sup> higher versus the GMM approach. These findings likely reflect the comparatively low density of observations at high-velocities in competitive team sports (Rampinini et al. 2007; Gregson et al. 2010; Datson et al. 2017), and question the utility of the *k*-means algorithm for prescribing velocity bins in athlete movement tracking.

Spectral Clustering was considered more appropriate for velocity data considering that the structure of the individual clusters (zones) is not suitably described by the center and dispersion of the complete data-set. The Spectral Clustering thresholds derived in the analysis were much larger than both *k*-means and GMM and provided thresholds within the range of those adopted in current industry practice. Specifically, the transition velocity to HVR was similar to that recommended for female players by Bradley and Vescovi (Bradley and Vescovi 2015), but entry points into VHVR and SPR zones resided between these recommendations and the generic thresholds commonly used for male external load data in football. However, application of the newly proposed velocity thresholds had a meaningful impact upon the distance-covered metrics, with *very-likely small* to *likely very-large* effect sizes observed when compared to Generic<sub>BV</sub> and Generic<sub>male</sub> (see

Table 4). Increasing the smoothing parameter increased the number of valid partitionings (as intended), but while slightly reducing the HSVR and VHSR entry thresholds, no meaningful change in the distances covered were observed. This may suggest that the Spectral Clustering method is robust to the choice of smoothing parameter. However, with no smoothing ( $\beta = 0$ ), 28.6% of the intervals were not useable, compared with 3.2% with  $\beta = 0.1$ . Accordingly, application of  $\beta = 0.1$  in this study modelled a larger data-set, which may derive more accurate thresholds. Further work maybe warranted to examine the impact of other smoothing parameters, although users are cautioned against over-smoothing raw Doppler velocity and acceleration data.

We acknowledge that the new thresholds learned in this study via data-mining techniques reflect the workload profiles of the sample, which are likely influenced by the positional role, the standard of opposition, match status and the squads tactical approach. Hence, the playing standard of women's football players should be considered by the load-monitoring practitioner prior to implementation. However, the modest sample size adopted in our analysis (227 match observations taken from 27 players) to some degree encapsulated the high-degree of between-match variation observed in football high-velocity running metrics (Gregson et al. 2010; Trewin et al. 2018). At the time of data-collection, GPS devices were not permitted for use in competitive tournaments or their qualification matches, and the current dataset was collected from preparatory fixtures. Accordingly, further work may be warranted to evaluate the new thresholds proposed in the current study using data sampled from multiple teams during competitive tournaments. Although the data-mining approaches employed in this study permit the determination of individual-specific zones (see Figure 1), development of new generic zone criteria were prioritized in this study given the high prevalence of generic velocity zones used to monitor player workloads in industry practice (Akenhead and Nassis 2016). While individualization of velocity zones maybe considered valuable to the interpretation and evaluation of player-specific work-loads, generic zones provide benchmarks for comparison both within and between teams.

In summary, the current study examined a range of data-mining techniques in an attempt to provide new time-motion analysis velocity thresholds for elite women's football players. This study adds new insights to the debate around the appropriate velocity thresholds to use for populations other than elite-male football players. We identified that *k*-means clustering and Gaussian mixture modeling were not appropriate for football given the limited instances in which players move at velocities associated with sprinting, which are often considered key physical performance indicators. A Spectral Clustering technique with application of a  $\beta = 0.1$  smoothing factor derived new thresholds featuring both logical validity and analysis rigor. Accordingly, we would recommend that examination of velocity data in elite women's football use 3.46 (12.5 km h<sup>-1</sup>), 5.29 (19.0 km·h<sup>-1</sup>), and 6.26 m·s<sup>-1</sup> (22.5 km·h<sup>-1</sup>), to denote entry into HVR, VHVR, and SPR generic categories, respectively. Similar analyses may be warranted to determine appropriate velocity zones for other sports and youth populations.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Laurence A. F. Park  <http://orcid.org/0000-0003-0201-4409>  
Ric Lovell  <http://orcid.org/0000-0001-5859-0267>

## References

- Akenhead R, Nassis GP. 2016. Training load and player monitoring in high-level football: current practice and perceptions. *Int J Sports Physiol Perform.* 11(5):587–593.
- Bangsbo J, Nørregaard L, Thorsø F. 1991. Activity profile of competition soccer. *Can J Sport Sci.* 16(2):110–116.
- Batterham AM, Hopkins WG. 2006. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform.* 1(1):50–57.
- Bradley PS, Dellal A, Mohr M, Castellano J, Wilkie A. 2014a. Gender differences in match performance characteristics of soccer players competing in the UEFA champions league. *Hum Mov Sci.* 33:159–171.
- Bradley PS, Lago-Penas C, Rey E. 2014b. Evaluation of the match performances of substitution players in elite soccer. *Int J Sports Physiol Perform.* 9(3):415–424.
- Bradley PS, Mascio MD, Peart D, Olsen P, Sheldon B. 2010. High-intensity activity profiles of elite soccer players at different performance levels. *J Strength Conditioning Res.* 24(9):2343–2351.
- Bradley PS, Vescovi JD. 2015. Velocity thresholds for women's soccer matches: sex specificity dictates high-speed running and sprinting thresholds - Female Athletes in Motion (FAiM). *Int J Sports Physiol Perform.* 10(1):112–116.
- Buchheit M, Mendez-Villanueva A, Simpson BM, Bourdon PC. 2010. Match running performance and fitness in youth soccer. *Int J Sports Med.* 31(11):818–825.
- Datson N, Drust B, Weston M, Jarman IH, Lisboa PJ, Gregson W. 2017. Match physical performance of elite female soccer players during international competition. *J Stren Cond.* 31(9):2379–2387.
- Dellal A, Hill-Haas S, Lago-Penas C, Chamari K. 2011. Small-sided games in soccer: amateur vs. professional players' physiological responses, physical, and technical activities. *J Strength Conditioning Res.* 25(9):2371–2381.
- di Prampero PE, Fusi S, Sepulcri L, Morin JB, Belli A, Antonutto G. 2005. Sprint running: a new energetic approach. *J Exp Biol.* 208(Pt 14):2809–2816.
- Di Salvo V, Baron R, Tschan H, Calderon Montero F, Bachi N, Pigozzi F. 2007. Performance characteristics according to playing position in elite soccer. *Int J Sports Med.* 28(3):222–227.
- Dwyer DB, Gabbett TJ. 2012. Global positioning system data analysis: velocity ranges and a new definition of sprinting for field sport athletes. *J Strength Conditioning Res.* 26(3):818–824.
- Gregson W, Drust B, Atkinson G, Salvo VD. 2010. Match-to-match variability of high-speed activities in premier league soccer. *Int J Sports Med.* 31(4):237–242.
- Harley JA, Barnes CA, Portas M, Lovell R, Barrett S, Paul D, Weston M. 2010. Motion analysis of match-play in elite U12 to U16 age-group soccer players. *J Sport Sci.* 28(13):1391–1397.
- Hopkins WG. 2007. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. *Sportscience.* 11:16–20. Retrieved from [sports.org/2007/wghinf.htm](http://sports.org/2007/wghinf.htm).
- Hunter F, Bray J, Towilson C, Smith M, Barrett S, Madden J, Abt G, Lovell R. 2015. Individualisation of time-motion analysis: a method comparison and case report series. *Int J Sports Med.* 36(1):41–48.
- Jennings D, Cormack S, Coutts AJ, Boyd L, Aughey RJ. 2010. The validity and reliability of GPS units for measuring distance in team sport specific running patterns. *Int J Sports Physiol Perform.* 5(3):328–341.
- Lovell R, Abt G. 2013. Individualization of time-motion analysis: a case-cohort example. *Int J Sports Physiol Perform.* 8(4):456–458.



- Malone JJ, Lovell R, Varley MC, Coutts AJ. 2017. Unpacking the black box: applications and considerations for using GPS devices in sport. *Int J Sports Physiol Perform.* 12(Suppl 2):S218–S226.
- Mohr M, Krstrup P, Bangsbo J. 2003. Match performance of high-standard soccer players with special reference to development of fatigue. *J Sport Sci.* 21(7):519–528.
- Mujika I, Santisteban J, Impellizzeri FM, Castagna C. 2009. Fitness determinants of success in men's and women's football. *J Sports Sci.* 27(2):107–114.
- Nassis GP, Brito J, Dvorak J, Chalabi H, Racinais S. 2015. The association of environmental heat stress with performance: analysis of the 2014 FIFA World Cup Brazil. *Br J Sports Med.* 49(9):609–613.
- Park LAF, Bezdek JC, Leckie C, Kotagiri R, Bailey J, Palaniswami M. 2016. Visual assessment of clustering tendency for incomplete data. *IEEE Trans Knowl Data Eng.* 28(12):3409–3422.
- Park LAF, Kotagiri R. 2011. Multiresolution web link analysis using generalized link relations. *IEEE Trans Knowl Data Eng.* 23(11):1691–1703.
- Park LAF, Leckie CA, Ramamohanarao K, Bezdek JC. 2009. Adapting spectral co-clustering to documents and terms using latent semantic analysis. In: Nicholson A, Li X, editors. *AI 2009: advances in Artificial Intelligence*. Vol. 5866. Berlin: Springer; p. 301–311.
- Park LAF, Simoff S (2013). Power walk (pp. 50–57). Presented at the 18th Australasian document computing symposium, new york, New York (USA): ACM Press.
- Rampinini E, Coutts AJ, Castagna C, Sassi R, Impellizzeri FM. 2007. Variation in top level soccer match performance. *Int J Sports Med.* 28(12):1018–1024.
- Randers MB, Mujika I, Hewitt A, Santisteban J, Bischoff R, Solano R, Zubillaga A, Peltola E, Krstrup P, Mohr M. 2010. Application of four different football match analysis systems: a comparative study. *J Sports Sci.* 28(2):171–182.
- Scott D, Lovell R. 2018. Individualisation of speed thresholds does not enhance the dose-response determination in football training. *J Sports Sci.* 36(13): 1523–1532.
- Scott MTU, Scott TJ, Kelly VG. 2016. The validity and reliability of global positioning systems in team sport: A brief review. *J Strength Conditioning Res.* 30(5):1470–1490.
- Siegle M, Lames M. 2010. The relation between movement velocity and movement pattern in elite soccer. *Int J Perform Anal Sport.* 10(3):270–278.
- Sweeting AJ, Aughey RJ, Cormack SJ, Morgan S. 2017a. Discovering frequently recurring movement sequences in team-sport athlete spatio-temporal data. *J Sports Sci.* 35(24): 2439–2445.
- Sweeting AJ, Cormack SJ, Morgan S, Aughey RJ. 2017b. When is a sprint a sprint? A review of the analysis of team-sport athlete activity profile. *Front Physiol.* 8:432.
- Trewin J, Meylan C, Varley MC, Cronin J. 2018. The match-to-match variation of match-running in elite female soccer. *J Sci Med Sport.* 21(2):196–201.