Some Issues in Distance Construction for Football Players Performance Data

Serhat Emre Akhanli and Christian Hennig

Abstract For mapping football (soccer) player information by using multidimensional scaling, and for clustering football players, we construct a distance measure based on players' performance data. The variables are of mixed type, but the main focus of this paper is how count variables are treated when defining a proper distance measure between players (e.g., top and lower level variables). The distance construction involves four steps: 1) representation , 2) transformation, 3) standardisation, 4) variable weighting. Several distance measures are discussed in terms of how well they match the interpretation of distance and similarity in the application of interest, with a focus on comparing Aitchison and Manhattan distance for variables giving percentage compositions. Preliminary outcomes of multidimensional scaling and clustering are shown.

Serhat Emre Akhanli
Department of Statistical Science, University College London Gower St, London WC1E 6BT, United Kingdom,
✉ serhat.akhanli.14@ucl.ac.uk

Christian Hennig
Department of Statistical Science, University College London Gower St, London WC1E 6BT, United Kingdom
✉ c.hennig@ucl.ac.uk

ARCHIVES OF DATA SCIENCE (ONLINE FIRST) DOI 10.5445/KSP/1000058749/09
KIT SCIENTIFIC PUBLISHING ISSN 2363-9881
Vol. 2, No. 1, 2017

1 Introduction

The wide range of interest in football stimulated much scientific interest, especially in statistics. High quality data are available on players' performance. In football, performance assessment has typically been conducted to predict players' abilities, to rate players' performances, to enhance their physical performance or to explain a team's success, see for example Mohr et al (2003); Di Salvo et al (2007); McHale et al (2012); Oberstone (2009).

Here we present a new idea to map football player information in order to explore their similarity structure. This type of information can be useful for football scouts and managers when assessing players, and also journalists and football fans will be interested. For instance, football scouts and managers try to find talented players that have certain characteristics to fit into their team and the system, and our map could help them to locate such players.

The data available to us are quite complex with many types of variables that may need some individual treatment, which cannot all be done in a single paper. The main focus of this paper is on the treatment of count variables in the construction of a distance, particularly in the situation in which there are top level count variables such as number of shots and lower level variables decomposing the top level variables into sub-categories, for example according to accuracy (on target, off target, blocked). The paper is meant to discuss some exemplary issues rather than to define the processing of all variables in full.

Note that clustering and mapping are unsupervised so that decisions cannot be made by optimising cross-validated prediction quality. We follow the principles for distance construction as explained in Hennig and Hausdorf (2006); Hennig and Liao (2013), according to which distances need to be constructed in such a way that they match as closely as possible the "interpretative distance" in the given application, i.e., how similar or different the objects are in terms of subject matter knowledge regarding the use of the distance, the resulting map, or the resulting clustering.

Section 2 introduces the football players data set. In Sect. 3, we give a general overview of steps required for the pre-processing of the variables in order to construct a distance, and summarise what was done in these steps for the football players' performance data. Section 4 treats the aggregation of the different variables with a focus on lower level compositional variables. In Sect. 5, results from multidimensional scaling and distance-based clustering are presented, which are currently preliminary and sketchy; improving on these is a topic for further research.

2 Football Players Dataset

The data set was obtained from the website `http://www.whoscored.com`. The data set contains 3152 football players characterized by 107 variables (as this is work in progress, we only present results from a subset of players as described below). The players are collected covering 8 major leagues (England, Spain, Italy, Germany, France, Russia, Netherlands, Turkey) based on the 2014-2015 football season. The data set consists of the players who have appeared at least in one game during the season. Goalkeepers have completely different characteristics from outfield players and were therefore excluded from our analysis. Variables are of mixed type, containing binary, count and continuous information. The variables can be grouped as follows:

Team and league (ratio scale number): League and team ranking score based on the information on the UEFA website, and team points from the ranking table of each league.

Position variables (binary): 15 variables indicating possible positions on which a player can play and has played.

Characteristic variables (ratio scale numbers): Age, height, weight,

Appearance variables (ratio scale numbers): Number of appearances of teams and players, and players number of minutes played.

Count variables (top level): Interceptions, fouls, offsides, clearances, unsuccessful touch, dispossess, cards, etc.

Count variables (lower level): Subdivision of some top level count variables as shown in Table 1.

3 Variable pre-processing

Data should be processed in such a way that the resulting distance between observations matches how distance is interpreted in the application of interest, see

Table 1: Top and lower level count variables

TOP LEVEL	*LOWER LEVEL*
SHOT	*Zone:* Out of box, six yard box, penalty area *Situation:* Open play, counter, set piece, penalty taken *Body part:* Left foot, right foot, header, other *Accuracy:* On target, off target, blocked
GOAL	*Zone:* Out of box, six yard box, penalty area *Situation:* Open play, counter, set piece, penalty taken *Body part:* Left foot, right foot, header, other
PASS	*Length:* AccLP, InAccLP, AccSP, InAccSP *Type:* AccCr, InAccCr, AccCrn, InAccCrn, AccFrk, InAccFrk
KEY PASS	*Length:* Long, short *Type:* Cross, corner, free kick, through ball, throw-in, other
ASSIST	*Type:* Cross, corner, free kick, through ball, throw-in, other
BLOCK	Pass blocked, cross blocked, shot blocked
TACKLE	Tackles, dribble past
AERIAL	Aerial won, aerial lost
DRIBBLE	Dribble won, dribble lost

*Acc: Accurate, *InAcc: Inaccurate
*LP: Long pass, *SP: Short pass, *Cr: Cross, *Crn: Corner, *Frk: Free kick

Hennig and Hausdorf (2006). The resulting dissimilarities between objects may strongly depend on transformation, standardization, etc., which makes variable pre-processing very important. Different ways of data pre-processing are not objectively “right” or “wrong”; they implicitly construct different interpretations of the data. We distinguish four key pre-processing steps:

1. Representation.
2. Transformation.
3. Standardisation.
4. Weighting.

3.1 Representation

This is about how to represent the relevant information in the variables, potentially defining new variables summarising or framing information in better ways. The count variables in the data set can be classified into two different categories: a) how many times the players perform an action overall (top level), b) within the action what compositions they have (lower level).

3.1.1 Top level count variables

A representation issue here is that in order to characterise players, counts of actions such as shots, blocks etc. should be used relative to the period of time the player played. A game of football lasts for 90 minutes, so we represent the counts as “per 90 minutes”:

$$y_{ij} = \frac{x_{ij}}{m_i/90} = 90 \times \frac{x_{ij}}{m_i}, \tag{1}$$

where x_{ij} is the j^{th} count variable of player i, m_i is the number of minutes played by player i.

3.1.2 Lower level count variables

Suppose that a player has 2.0 shots per 90 minutes, and the shots per zone are out of box: 0.4, penalty area: 1.3, six yard box: 0.3. When computing the distance between this player and another player, two different aspects of the players’ characteristics are captured in these data, namely how often a player shoots, and how the shots distribute over the zones. If the data were used in the raw form given above, players with a big difference in the top level variable “shots” would also differ strongly regarding the lower level variable “shot zone”, and the overall distance would be dominated by the top level variable with the information on the zonal distribution being largely lost. In order to separate the different aspects of interest, the lower level count variables are transformed to percentages, i.e., 0.2, 0.65 and 0.15 for out of box, penalty area, six yard box above, whereas the top level count is taken as per 90 minutes count as defined above (before transformation, see below).

Percentage variables can be represented as proportional total and/or success rates. For example, shot and goal are top level count variables that contain common sub-categories (zone, situation, body part). Goal is essentially the successful completion of a shot, so that the sub-variables of goal can be treated as success rate of shot in the respective category as well as as composition of total goals. Both are of interest for characterising the players in different ways,

Table 2: Representation of lower level count variables

Variables **(Include sub-categories)**	**Proportional total** **(standardised by)**	**Success rate** **(standardised by)**
Block	Total Blocks	✘
Tackle, Aerial, Dribble	✘	Total tackles, total aerials, and total dribbles
Shot (4 sub-categories)	Total shots	✘
Goal (4 sub-categories)	Total goals	Shot count in different sub-categories, and total shots for overall success rate
Pass (2 sub-categories)	Total passes	Pass count in different sub-categories, and total passes for overall success rate
Key pass (2 sub-categories)	Total key passes	✘
Assist	Total assists	Key pass count in different sub-categories, and total assists for overall success rate

and therefore we will use both representations in some cases. Table 2 shows where this was applied.

3.2 Transformation

Variables are not always related to “interpretative distance” in a linear way, and transformation should be applied in order to match interpretative distances by the effective differences on the transformed variables.

The top level count variables have more or less skew distributions; for example, many players, particularly defenders, shoot very rarely during a game, and a few forward players may be responsible for the majority of shots. On the other hand, most blocks come from a few defenders, whereas most players block rarely. This means that there may be large absolute differences between players that shoot or block often, whereas differences at the low end will be low; but the interpretative distance between two players with large but fairly different numbers of blocks and shots is not that large, compared with the difference between, for example, a player who never shoots and one who occasionally but rarely shoots.

This suggests a non-linear concave transformation such as logarithm or square root for these variables, which effectively shrinks the difference between large values relative to the difference between smaller values. The exact choice of the transformation will be discussed elsewhere; a guiding principle can be the stabilisation of the variation of these variables as function of their values between different seasons for the same player.

No transformation is applied to compositional percentages, because a difference of, say, 0.05, has the same meaning in each category regardless of whether this is a difference between high or low percentages.

3.3 Standardisation

Whereas transformation deals with relative differences between players on the same variable, standardisation and weighting are about calibrating the impact of the different variables against each other. Usually, weighting and standardisation both involve multiplying a variable with a constant, but they have different meanings. Standardisation is about making the measurements of the different variables comparable in size, whereas weighting is about giving the variables an impact on the overall distance that corresponds to their subject-matter importance.

We standardise transformed top level count variables to unit variance, see Hennig and Liao (2013) for a discussion of this. For the lower level percentages, we standardise by dividing by the pooled average L_1-distance from the median. We pool this over all categories belonging to the same composition of lower level variables. This means that all category variables of the same composition are standardised by the same value, regardless of their individual relative variances. The reason for this is that a certain difference in percentages between two players has the same meaning in each category, which does not depend on the individual variance of the category variable.

3.4 Weighting

One aspect of variable weighting here is that in case that there are one or more lower level compositions of a top level variable, the top level variable is transformed and standardised individually whereas the categories of the composition are standardised together. This reflects the fact that the top level count and the lower level distribution represent distinct aspects of a player's characteristics, and on this basis we assign the same weight to the top level variable as to the whole vector of compositional variables, e.g., a weight of one for transformed shot counts is matched by a weight of $1/3$ for each of the zone variables “out of the box”, “six yard box”, “penalty area”.

The percentage variables of the same composition are linearly dependent and are therefore correlated with each other; k percentage variables do not represent k independent parts of information. Variable selection and dimension reduction are very popular to deal with this. However, in distance construction, the problem is appropriately dealt with using weighting. There is no advantage in using, for example, only two variables out of “out of the box”, “six yard box”, “penalty area” and weight them by $1/2$ each; using all three means that they are treated symmetrically for the construction of the distance, as is appropriate. At the same time down-weighting avoids that the redundant information dominates the overall distance.

In case that a top level count variable is zero for a player, the percentage variables are missing. In this situation, for overall distance computation between such a player and another player, the composition variables can be assigned weight zero and the weight that is normally on a top level variable and its low level variables combined can be assigned to the top level variable.

A potential approach for computing distances between count and composition variables that could be seen as relevant here is the χ^2-distance (Greenacre, 2007), which implicitly weights counts by the inverse of marginal (variable and observation) totals. We will not use this approach here. For the top level variables, computing observation totals by summing up counts from different variables is not appropriate here; these do not reflect appropriately the relative importance of the variables. For the low level compositions, see Section 4.2.2.

4 Aggregation of variables

There are different well-known ways of aggregating variables in order to define a distance measure such as the Euclidean or Manhattan distance. The same principle as before, “matching interpretative distance”, applies here as well.

4.1 Top level variables

There are different types of variables in this data set (the position variables are treated in a non-Euclidean way, which will be explained elsewhere), and therefore we decided against using Euclidean aggregation, which implicitly treats the variables as if they are in a joint Euclidean space, and which weights larger differences on individual variables up when comparing two players. Instead, we aggregate variables by summing the individual distances up, i.e., following the principle for the Manhattan distance, as also used by Gower (1971). This means that distances on all variables are treated in the same way regardless of the size of the difference.

4.2 Lower level compositional percentage variables

4.2.1 Aitchison's theory for compositional data

Our percentage variables are compositional data in the sense of Aitchison (1986), who set up an axiomatic theory for the analysis of compositional data. We will argue here that for our application for the compositional percentage data the simple Manhattan distance is more appropriate than what Aitchison proposed specifically for compositional data, which means that the principle of matching interpretative distance in distance construction can be in conflict, depending on the application, with a pure mathematical axiomatic approach.

$\mathbf{x} = [x_1, x_2, ..., x_D]$ is a D-part composition when all its components are strictly positive real numbers and carry only relative information. The sample space of compositional data is the D dimensional positive simplex,

$$S^D = \{[x_1, x_2, ..., x_D] | x_i > 0, i = 1, 2, ..., D; \sum_{i=1}^{D} x_i = c\}, \qquad c = 1 \text{ here.} \tag{2}$$

Aitchison (1992) proposed four requirements for any distance d for compositional data:

1. **Scale invariance:** For any positive real value $\lambda \in \Re_+$: $d(\lambda\mathbf{x}, \lambda\mathbf{y}) = d(\mathbf{x}, \mathbf{y})$.
2. **Permutation invariance:** Reordering the parts of the composition should not change the distance.
3. **Perturbation invariance:** Let $\mathbf{x}, \mathbf{y} \in S^D$ and $q = (q_1, q_2, \dots q_d)$, $q \in \Re^D_+$. Then,
$$d(q \otimes \mathbf{x}, q \otimes \mathbf{y}) = d(\mathbf{x}, \mathbf{y}) \; for\ every\ perturbation\ q, \tag{3}$$
where "$\otimes$" stands for component-wise multiplication.

4. **Sub-compositional coherence:** For a sub-composition $\mathbf{x_s}, \mathbf{y_s}$ of $\mathbf{x}, \mathbf{y} \in S^D$, namely a subset of the components of **x,y**:

$$d(\mathbf{x}, \mathbf{y}) \geq d(\mathbf{x_s}, \mathbf{y_s}). \tag{4}$$

Aitchison then showed that the “Aitchison distance” d_a is one of few distance measures to fulfill the axioms (Aitchison, 1992):

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{i=1}^{D} \{\ln \frac{x_i}{g_m(\mathbf{x})} - \ln \frac{y_i}{g_m(\mathbf{y})}\}^2}, \tag{5}$$

where $g(\cdot) = (\prod_{i=1}^{D} x_i)^{1/D}$ (Geometric mean of the compositions).

4.2.2 Manhattan distance vs Aitchison distance

In order to compare the Manhattan and the Aitchison distance, we first discuss the Manhattan distance regarding the four axioms.

- The Manhattan distance does not fulfill **scale invariance**; if both compositions are multiplied by λ, the Manhattan distance is multiplied by λ. This, however, is irrelevant here, because we are interested in percentages only that sum up to 1, so multiplication does not happen.
- The Manhattan distance is **permutation invariant**.
- The Manhattan distance is not **perturbation invariant**, but as was the case for scale invariance, this is irrelevant here, because the percentages are relative counts and the operation of multiplying different categories in the same composition with different constants is not meaningful in this application.
- The Manhattan distance is not **sub-compositional coherent**, but once more this is not relevant, because in this application it is not meaningful to compare the values of the compositional distance with those from sub-compositions.

Aitchison’s axioms were proposed as general principles for compositional data, but in fact the axioms were motivated by specific applications with specific characteristics, which mostly do not apply here. Note that the Aitchison distance is not defined when any of x_i’s or y_i’s are zero, because then the term inside the logarithmic function becomes $0/0$. Aitchison (1986) and Martın-Fernandez et al (2011) proposed some modification techniques for zero proportions.

Furthermore, the Aitchison distance can be problematic for small percentages. For football players, the Aitchison distance does not seem suitable for matching “interpretative distance”. We demonstrate this using three popular players from the data set, namely James Rodriguez (JR), Alexis Sanchez (AS) and Cesc Fabregas (CF), and the “Block” action.

Table 3: Percentage variables in block action for three selected players

Players	**Shot blocked**	**Cross blocked**	**Pass blocked**
James Rodriguez (JR)	0.03	0.03	0.94
Alexis Sanchez (AS)	0.00 (≈ 0)	0.04	0.96
Cesc Fabregas (CF)	0.09	0.05	0.86

Table 4: Distances of block percentages for the three selected players

Distance	**JR-AS**	**JR-CF**	**AS-CF**
Manhattan	0.06	0.16	0.20
Aitchison	26.69	0.84	27.42

Percentages and distances are presented in Table 3 and Table 4. AS has a very small proportion (≈ 0 but nonzero) in the sub-variable of "Shot blocked". The Aitchison distance between JR and AS is quite large, whereas it is not large between JR and CF. But actually JR and AS are quite similar players according to the data; both block almost exclusively passes and hardly any shots or crosses. CF blocks substantially more shots and some more crosses than both others and therefore the two distances between CF and both JR and AS should be bigger than that between JR and AS, which is what the Manhattan distance delivers. The Manhattan distance treats absolute differences between percentages in the same way regardless of the size of the percentages between which these differences occur. The Aitchison distance is dominated by differences between small percentages in an inappropriate manner.

In addition, the general principle is that differences on different variables should be treated the same. We actually want to have the resulting distances between percentages to count in the same way regardless of which part of the compositions they are. This argument can be made formal by the following theory:

Definition 1. Let $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ be a D-part composition from the data set X, $i = 1, 2, \dots, n$, with the following assumptions

i $\sum_{k=1}^{D} x_{ik} = 1$,

ii $0 \leq x_{ik} \leq 1$,

iii $D > 2$,

and let s_k, $k = 1, 2, \dots, D$, be a standardised constant which may or may not depend on k for all $s_k > 0$. Then, consider the following distances:

1. Standardised Euclidean distance:

$$d_E(x_i, x_j) = \sqrt{\sum_{k=1}^{D} \left\{ \frac{x_{ik}}{s_k} - \frac{x_{jk}}{s_k} \right\}^2}, \tag{6}$$

Note that the $\chi^2 - distance$ as mentioned in Section 3.4 is a standardised Euclidean distances with s_k depending on k in general.

2. Standardised Manhattan distance:

$$d_M(x_i, x_j) = \sum_{k=1}^{D} \left| \frac{x_{ik}}{s_k} - \frac{x_{jk}}{s_k} \right|, \tag{7}$$

3. Aitchison distance:

$$d_A(x_i, x_j) = \sqrt{\sum_{k=1}^{D} \left\{ \log\left(\frac{x_{ik}}{g(x_i)} \right) - \log\left(\frac{x_{jk}}{g(x_j)} \right) \right\}^2}, \tag{8}$$

where $g(x_i) = \left(\prod_{k=1}^{D} x_{ik} \right)^{1/D}$.

Axiom 1. Let $x_1 = (x_{11}, x_{12}, \dots, x_{1D})$, $x_1^{(1)} = (x_{11} + \varepsilon, x_{12} - \frac{\varepsilon}{D-1}, \dots, x_{1D} - \frac{\varepsilon}{D-1})$, $x_1^{(q)} = (x_{11} - \frac{\varepsilon}{D-1}, \dots, x_{1q} + \varepsilon, \dots, x_{1D} - \frac{\varepsilon}{D-1})$ be D-part compositions. Assume that $x_1, x_1^{(1)}, x_1^{(q)} \in X$, $0 < \varepsilon \leq 1$, and $0 \leq x_{1k}^{(t)} \leq 1$ for all t. Hence, the general distance satisfies the following equation;

$$d(x_1, x_1^{(1)}) = d(x_1, x_1^{(q)}), \tag{9}$$

Theorem 1. *Equation 9 does hold for d_E and d_M if and only if $s = s_k$ $\forall k$, and does not hold for d_A ($x_{1k}^{(t)} \neq 0$) in general.*

Proof. Standardised Euclidean distance:

$$\begin{aligned}
d_E(x_1, x_1^{(1)})^2 - d_E(x_1, x_1^{(q)})^2 &= \sum_{k=1}^{D} \left\{ \frac{x_{1k}}{s_k} - \frac{x_{1k}^{(1)}}{s_k} \right\}^2 - \sum_{k=1}^{D} \left\{ \frac{x_{1k}}{s_k} - \frac{x_{1k}^{(q)}}{s_k} \right\}^2 \\
&= \left(\frac{\varepsilon^2}{s_1^2} + \frac{\varepsilon^2}{(D-1)^2} \sum_{k=2}^{D} \frac{1}{s_k^2} \right) - \left(\frac{\varepsilon^2}{s_q^2} + \frac{\varepsilon^2}{(D-1)^2} \sum_{\substack{k=1 \\ k \neq q}}^{D} \frac{1}{s_k^2} \right) \\
&= \varepsilon^2 \left[\frac{1}{s_1^2} - \frac{1}{s_q^2} + \frac{1}{(D-1)^2} \left(\frac{1}{s_1^2} - \frac{1}{s_q^2} \right) \right] \\
&= \varepsilon^2 \left(\frac{1}{s_1^2} - \frac{1}{s_q^2} \right) \left(1 - \frac{1}{(D-1)^2} \right) \\
&= 0 \iff \quad s_1 = s_q.
\end{aligned}$$

If this is satisfied $\forall q$, then $s = s_k \; \forall k$.

Standardised Manhattan distance:

$$\begin{aligned}
d_M(x_1, x_1^{(1)}) - d_M(x_1, x_1^{(q)}) &= \sum_{k=1}^{D} \left| \frac{x_{1k}}{s_k} - \frac{x_{1k}^{(1)}}{s_k} \right| - \sum_{k=1}^{D} \left| \frac{x_{1k}}{s_k} - \frac{x_{1k}^{(q)}}{s_k} \right| \\
&= \left(\frac{\varepsilon}{s_1} + \frac{\varepsilon}{D-1} \sum_{k=2}^{D} \frac{1}{s_k} \right) - \left(\frac{\varepsilon}{s_q} + \frac{\varepsilon}{D-1} \sum_{\substack{k=1 \\ k \neq q}}^{D} \frac{1}{s_k} \right) \\
&= \varepsilon \left[\frac{1}{s_1} - \frac{1}{s_q} + \frac{1}{D-1} \left(\frac{1}{s_1} - \frac{1}{s_q} \right) \right] \\
&= \varepsilon \left(\frac{1}{s_1} - \frac{1}{s_q} \right) \left(1 - \frac{1}{D-1} \right) \\
&= 0 \iff \quad s_1 = s_q.
\end{aligned}$$

If this is satisfied $\forall q$, then $s = s_k \; \forall k$.
Aitchison distance: The proof will be provided by counter examples. Table 5 proves that the Aitchison distance does only fulfil Equation 9 for $x_{11} = x_{1q}$.

Table 5: Counter examples for the proof of the Aitchison distance

Compositions, where $\varepsilon = 0.15$	**For** $x_{11} \neq x_{1q}$	**For** $x_{11} = x_{1q}$
x_1	$(0.40, 0.30, 0.20, 0.10)$	$(0.25, 0.30, 0.25, 0.20)$
$x_1^{(1)}$	$(0.55, 0.25, 0.15, 0.05)$	$(0.40, 0.25, 0.20, 0.15)$
$x_1^{(3)}$	$(0.35, 0.25, 0.35, 0.05)$	$(0.20, 0.25, 0.40, 0.15)$
$d_A(x_1, x_1^{(1)}) - d_A(x_1, x_1^{(3)})$	-0.0368	0.0000

Remark: For $D = 2$, all reasonable standardisations are the same for both variables, since $x_{i1} = 1 - x_{iD}$.

Remark: In Correspondence Analysis (Greenacre, 2007), another central axiom is the "principle of distributional equivalence", which states that if two columns (resp., two rows) of a contingency table have the same relative values, then merging them does not affect the dissimilarities between rows (resp., columns). We are here only concerned with dissimilarities between players, not with dissimilarities between variables. For dissimilarities between players, distributional equivalence holds when using a standardised Manhattan distance with s_k chosen independently of k pooling average L_1-variable-wise distances from the median, because when merging two variables $\mathbf{x}$ and $\mathbf{y} = c\mathbf{x}$, these simply sum up.

5 Results

In order to show a preliminary analysis for a test subset of players based on the constructed distance including all variables, we show two Multidimensional Scaling (MDS) maps here, namely a) an MDS of the distances constructed as explained here (actually due to the lack of space, some details were omitted, e.g., the handling of the position variables), and b) an MDS of plain standardised Euclidean distances for all variables. There are various MDS techniques, see, e.g., Borg et al (2013). We use ratio MDS here, computed by the R-package "smacof" (de Leeuw and Mair, 2009). This means that the Euclidean distances on the map approximate a normalised version of the original dissimilarities in the sense of least squares. We adopt *The Partitioning Around Medoid* (PAM) clustering (Kaufman and Rousseeuw, 2009) with number of clusters $k = 6$.

Note that finding an optimal k and comparing different clustering methods is an issue for future work.

According to Figure 1, Ricardo Rodriguez (left back) and De Bruyne or Hazard (attacking midfielder) are quite different, but in the same cluster in plain Euclidean solution. Since both Rodriguez and Shaw play in the left back, they can be expected to be similar, which they are according to the distances constructed here, but in different clusters in plain Euclidean solution.

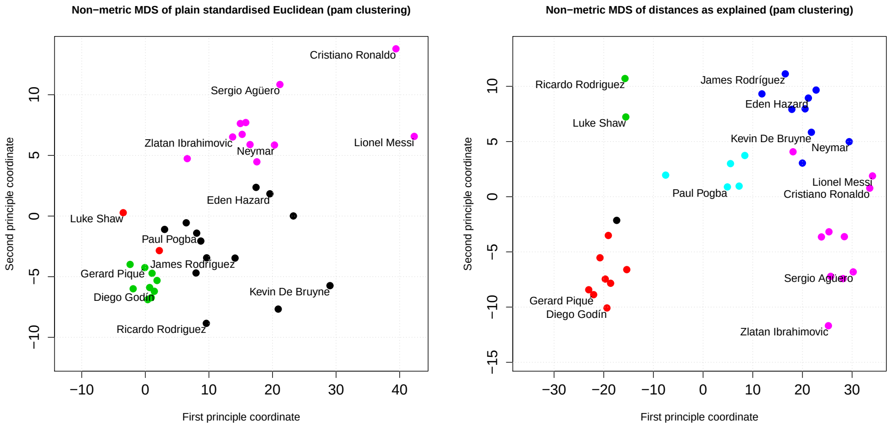

Fig. 1: Multidimensional scaling (MDS) and PAM clustering ($k = 6$) for test subset of players based on all variables.

The result implies that clustering and mapping multivariate data are strongly affected by pre-processing decisions such as the choice of variables, transformation, standardisation, weighting. The variety of options is huge, but the fundamental concept is to match the “interpretative dissimilarity” between objects as well as possible by the formal dissimilarity between objects. This is an issue involving subject-matter knowledge that cannot be decided by the data alone.

Acknowledgements The work of Christian Hennig was supported by EPSRC-grant EP/K033972/1.

References

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, Ltd., London

Aitchison J (1992) On criteria for measures of compositional difference. Mathematical Geology 24(4):365–379, DOI 10.1007/BF00891269

Borg I, Groenen PJ, Mair P (2013) Applied Multidimensional Scaling. Springer, Berlin, DOI 10.1007/978-3-642-31848-1

Di Salvo V, Baron R, Tschan H, Calderon Montero F, Bachl N, Pigozzi F (2007) Performance characteristics according to playing position in elite soccer. International Journal of Sports Medicine 28(3):222, DOI 10.1055/s-2006-924294

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27(4):857–874, DOI 10.2307/2528823, URL `http://www.jstor.org/stable/2528823`

Greenacre M (2007) Correspondence analysis in practice, 2nd edn. CRC Press, Boca Raton

Hennig C, Hausdorf B (2006) Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. In: Batagelj V, Bock HH, Ferligoj A, Ziberna A (eds) Data Science and Classification, Springer, Berlin, pp 29–38

Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics) 62(3):309–369, DOI 10.1111/j.1467-9876.2012.01066.x

Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis, vol 344. John Wiley & Sons, New York, DOI 10.1002/9780470316801

de Leeuw J, Mair P (2009) Multidimensional scaling using majorization: Smacof in R. Journal of Statistical Software 31(1):1–30, DOI 10.18637/jss.v031.i03, URL `https://www.jstatsoft.org/index.php/jss/article/view/v031i03`

Martın-Fernandez JA, Palarea-Albaladejo J, Olea RA (2011) Dealing with zeros, John Wiley & Sons, pp 43–58. DOI 10.1002/9781119976462.ch4

McHale IG, Scarf PA, Folker DE (2012) On the development of a soccer player performance rating system for the English Premier League. Interfaces 42(4):339–351, DOI 10.1287/inte.1110.0589

Mohr M, Krustrup P, Bangsbo J (2003) Match performance of high-standard soccer players with special reference to development of fatigue. Journal of sports sciences 21(7):519–528, DOI 10.1080/0264041031000071182, pMID: 12848386

Oberstone J (2009) Differentiating the top english premier league football clubs from the rest of the pack: Identifying the keys to success. Journal of Quantitative Analysis in Sports 5(3), DOI 10.2202/1559-0410.1183