All-round speed skating championships; towards a fair play using principal component analysis

Introduction

At an all-round speed skating tournament, both women and men skate four distances. For men, these are 500, 1.500, 5.000 and 10.000 meters. Women skate 500, 1.500, 3.000 and 5.000 meters. Each individual skater obtains points for each distance. The time at a specific distance is converted to a time at 500 meters. A difference of one second at the 500 meter needs to be compensated by 10 seconds at the 5.000 meter. The skater with the least points over four distances wins. The International Skating Union decided to replace the longest distance for men and women by a 1.000 meter. In this article, we show that this benefits skaters that are relatively fast at short distances. At the other hand, we show that long-distance skaters are currently in favour. We start by categorizing skaters using principal component analysis. Then we analyze the correlation between the points in a set-up and the type of skater. We propose two alternative set-ups and end with conclusions.

Using principal component analysis to categorize speed skaters

Principal component analysis is a statistical tool to obtain factors that explain variance in data. We apply this technique to the personal bests of 2.449 male speed skaters at six distances, using the database of Jakub Majerski. First, we normalize the data:

$$\widehat{P_{i,d}} = \frac{P_{i,d} - \frac{1}{n}\sum_{i=1}^{n} P_{i,d}}{\frac{1}{n-1}(P_{i,d} - \frac{1}{n}\sum_{i=1}^{n} P_{i,d})^2} \quad (1)$$

$P_{i,d}$= Personal best of speed skater $i \in \{1, \ldots, 2449\}$ at distance *d;*

$d \in \{1, \ldots, 6\}$ *(500, 1.000, 1.500, 3.000, 5.000 and 10.000 meters);*

$\widehat{P_{i,d}}$ = Normalized personal best.

Then, we apply singular value decomposition:

$$\widehat{P_{i,d}} = \sum_{j} U_{ij} \Sigma_{\mathrm{jj}} \mathrm{V}_{\mathrm{dj}} \quad (2)$$

U_{ij}= Interaction of skater *i* with principal component *j*;

Σ_{jj}= Principal component *j*;

V_{dj}= Interaction of personal best at distance *d* with principal component *j*.

Matrix $\boldsymbol{\Sigma}$ is diagonal, and the elements are ordered from large (explaining most of the variance) to small (explaining least of the variance). Matrix *V* contains vectors from which we can deduct how the times at specific distances are related to the principal components. Matrix *U* reflects the interplay between the skaters and the principal components.

Because we take into account six distances, we obtain six principal components (pc's). The first principal component reflects the level of the speed skater and explains 66% of the variance in the data. Some skaters are faster than others, regardless of distance. The second component categorizes skaters into sprinters and stayers and explains 24% of the variance. Sprinters are relatively fast at shorter distances, stayers are relatively fast at longer distances. In Figure 1 we can see the mapping of all 2.449 skaters onto these two factors. Next to that, the first two vectors of matrix *V* are shown.

The other factors all explain less than 5% of the variance, and reflect the fact whether skaters are relatively fast at intermediate distances (pc 3) or at specific distances (pc 4,5 and 6).

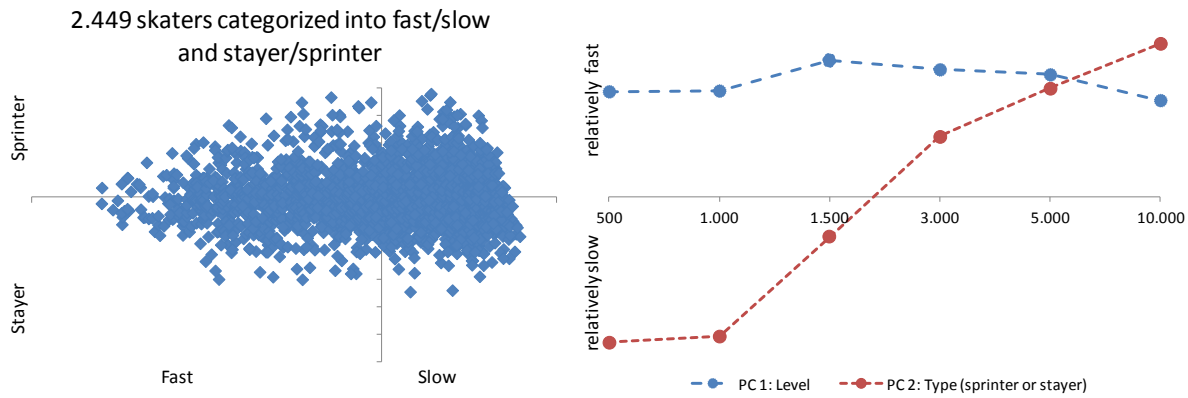

Figure 1: Categorizing 2249 male skaters by level and type, the first two principal components.

Is the current all-round tournament fair?

We would like to investigate whether both sprinters and stayers have equal chances in an all-round tournament. First, we look at the top 20 of the world, based on personal bests. In Figure 2 we see that in the current system, the top 20 is dominated by stayers. On the other hand, sprinters dominate in the new set-up. Shani Davis is on top of both lists. But if we look at Sven Kramer, we see that he is second best in the current system, but only at place 15 in the new set-up. A sprinter like Simon Kuipers, which is not in the current top 20, appears at place 5 in the new system. In Figure 3 we see scatter plots of points versus the first two principal components. We see a positive correlation between 'type' and points in the current set-up, and a negative correlation in the new set-up, which means stayers are indeed in favour in the current system, and sprinters are in favour in the new set-up. Table 1 shows the correlations.

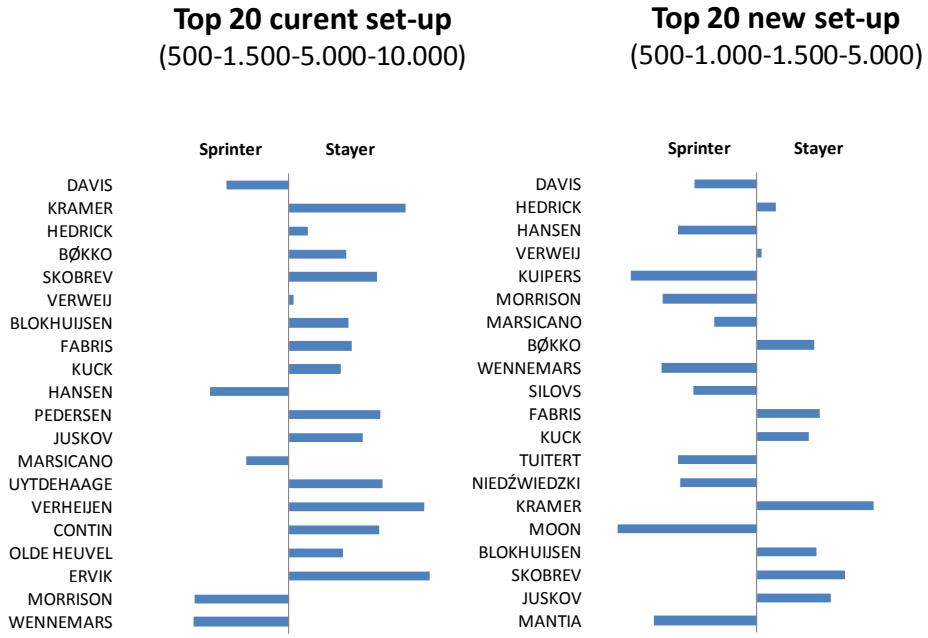

Figure 2: Top 20 in current and new system based on personal bests. Skaters are given a 'sprinter' or 'stayer' ranking based on percentiles of the second principal component.

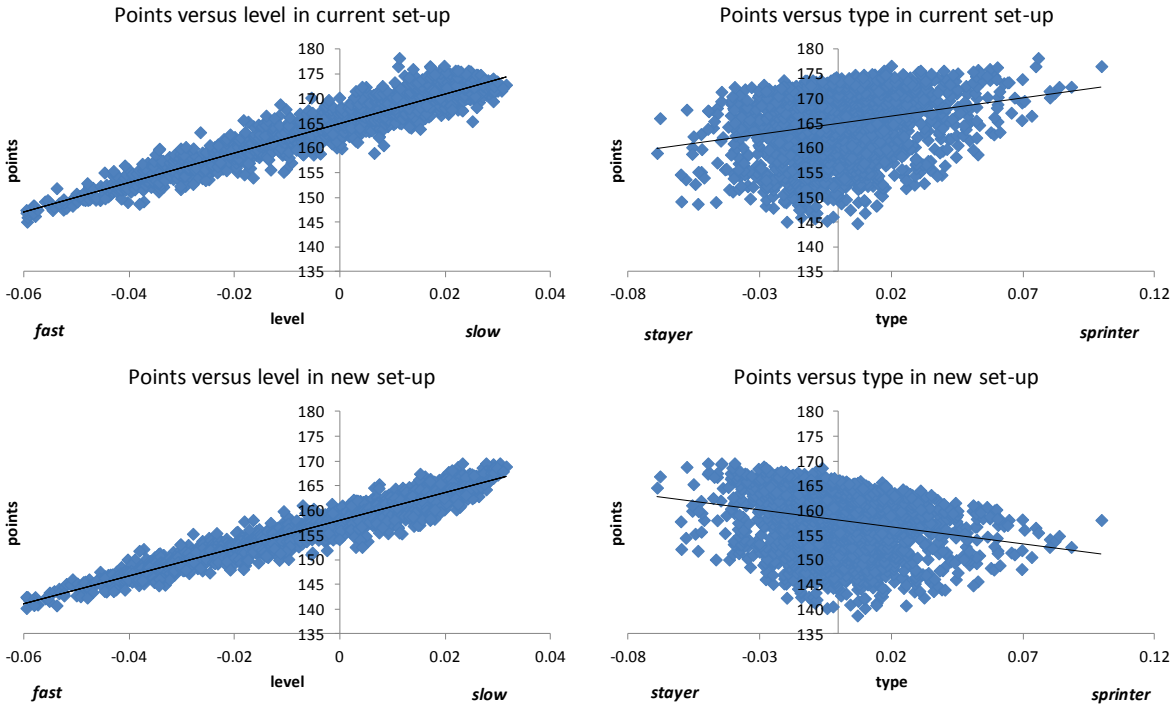

Figure 3: Scatter plot of points in the current set-up versus the first two principal components. A negative 'level' is a relatively fast skater. A positive 'type' represents a sprinter, a negative 'type' a stayer.

Towards a fair alternative

We would like to find an alternative tournament in which the level of the skater is most important, and neither sprinters nor stayers are in advantage at start. In the first alternative, we replace the 10.000 meter by a 3.000 meter instead of a 1.000 meter. In the second alterative, we do not replace it, and the skaters compete over three distances. In both alternatives, the correlation between the number of points and the level of the skater increases. And as we can see in Figure 4, in both set-ups, the correlation with the type of skater decreases in absolute terms. This is also reflected in the top 20 based on personal bests, which is shown in Figure 5.

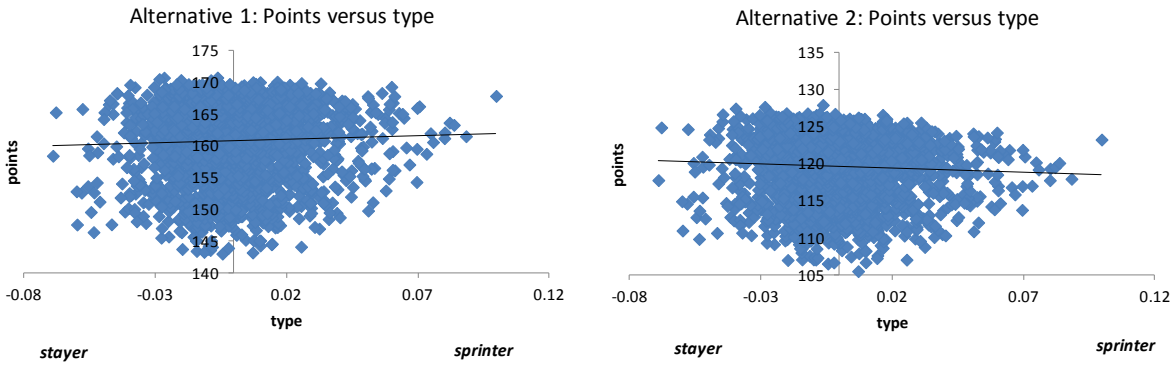

Figure 4: Scatter plot of points in the alternative set-up versus the second principal components 'type'.

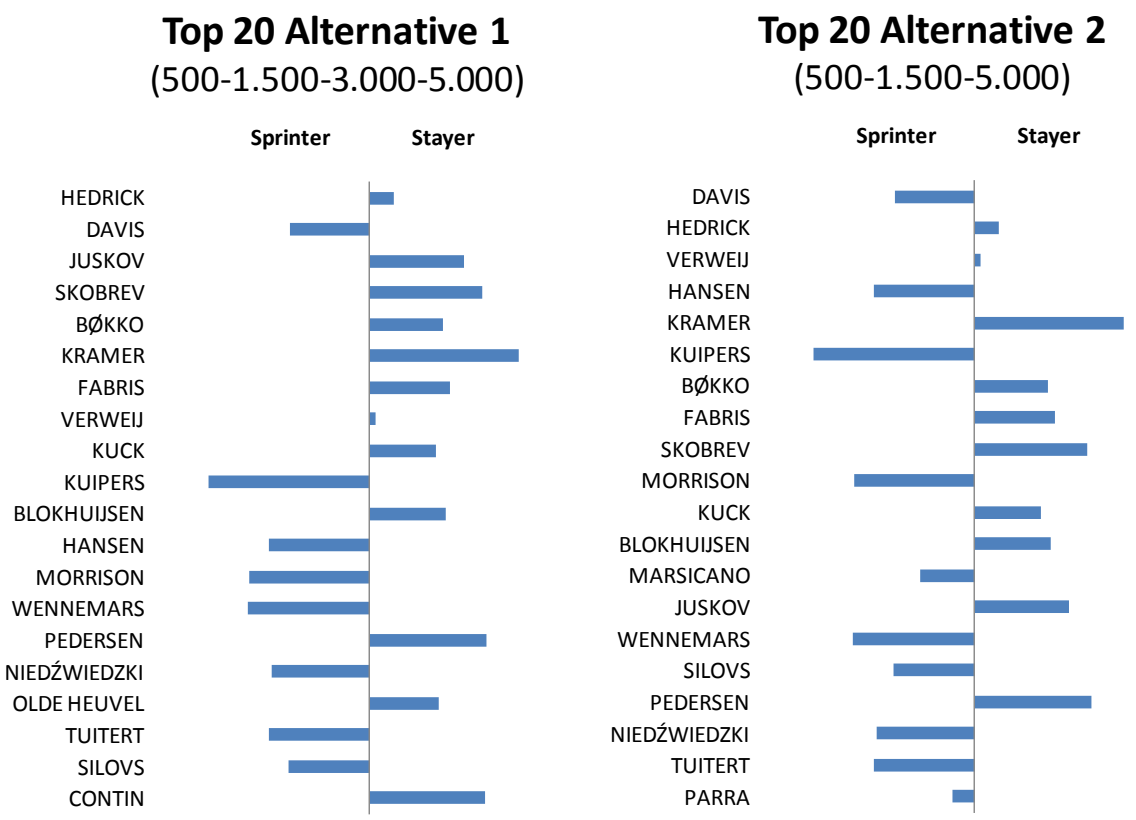

Figure 5: Top 20 in two alternative systems based on personal bests. The balance between stayers and sprinters increases compared to the current system.

	Current set-up (500, 1.500, 5.000, 10.000)	**New set-up** (500, 1.000, 1.500, 10.000)	**Alternative 1** (500, 1.000, 3.000, 5.000)	**Alternative 2** (500, 1.500, 5.000)
Level	96%	97%	99%	99%
Type	24%	-23%	4%	-5%

Table 1: Correlations between points in different set-ups, and the level and type of the speed skater. A positive correlation with the factor 'type' means that stayers are in favour.

Discussion of data

The database consists of skaters of whom at least one personal best lies below the threshold of the specific distance (40,00 1.20,00 2.03,00 4.18,00 7.30,00 16.00,00). Skaters are in the combined database of all six distances if the sum of the six personal bests, converted to 500 meter times, is less than 257 seconds. Therefore, skaters can be absent in the combined database for two reasons: if they do not have an official time at all six distances, or if the sum of their personal bests is too slow. Skaters are only in the combined dataset if they are able to skate a reasonably fast time at all six distances.

	500	**1.000**	**1.500**	**3.000**	**5.000**	**10.000**
Size dataset	9295	8850	8961	8400	7588	4264

Table 2: Overview of the number of skaters in the dataset of specific distances

The size of the 500 meter database is twice as large as the 10.000 meter database. In total, the combined dataset consists of 2449 skaters. In general, it is more common for a stayer to skate a short distance, and there are more sprinters that do not have set an official time at the long distances. Therefore, there are relatively more stayers in the database, compared to sprinters. However, note that almost half of the 10.000 meter skaters is not in the combined database, most of the times because their total time is too slow. In absolute terms, there are a bit more sprinters in the combined database (54% versus 46%).

Conclusions

Principal component analysis can be used to categorize a speed skater as a sprinter or a stayer. The current set-up of all-round tournaments benefits stayers. The new set-up, in which the 10.000 meter is replaced by the 1.000 meter, is in favour of sprinters. Principal component analysis can be used to find a fair alternative, in which the correlation with the level of the skater is higher, and the correlation with the type of skater is lower. We find two alternatives that satisfy this requirements: replace the 10.000 meter by a 3.000 meter, or do not replace it, and skate three distances only.

References

Jakub Majerski's speed skating database:
http://www.sskating.com/list.php?season=2013/2014&lm=m&dist=Big6