**ISSN** 1750-9823 (print)
**International Journal of Sports Science and Engineering**
Vol. 02 (2008) No. 02, pp. 123-128

# Large-scale Cluster Analysis of Elite Male and Female Swimmers'Race Patterns

Ilene CHEN[1,2 +], Ming-Yang CHEN[3], Craig JIN[1], Hong YAN[1,4]

[1] School of Electrical and Information Engineering, University of Sydney, Australia

[2] Swimming Australia, Australia

[3] Department of Electrical Engineering, Stanford University, USA

[4] Department of Electronic Engineering, City University of Hong Kong,Hong Kong

(*Received January10, 2008, accepted April 1, 2008*)

**Abstract.** The underlying theme of this paper is to show how large-scale cluster analysis can be applied to discover the best pacing strategy and the optimum combination of race components that will help athletes achieve their best performance at each particular race distance, which may not be observed easily otherwise. Taking into consideration the dependence of the repeated measurements made over athletes' entire race record times over each of the i-th laps, we report three main findings: Firstly, based on the resultant tree-like dendrograms from cluster analysis, we find that male and female swimmers may use similar patterns on stroke lengths in their races whereas the swimming speeds of males are generally somewhat faster. Secondly, we demonstrate a potential application of large-scale cluster analysis to find the optimum relationship between stroke length and swimming speed that will allow German national swimmer, Stefan Herbst, and French national swimmer, Amaury Leveaux to reproduce fast swims in the 200m freestyle race distance events. Thirdly, we demonstrate a potential application of large-scale cluster analysis to find the best pacing strategy that will allow British national swimmer, Paul Palmer, Australian national swimmer, Ian Thorpe, American national swimmer, Joshua Davis, Canadian national swimmer, Rick Say, Italian national swimmers, Emiliano Brembilla and Massimiliano Rosolino, to improve their times in the 200m freestyle race distance events.

**Key words:** large-scale cluster analysis, optimum relationship, swimming speed, stroke length, pacing strategy

## 1. INTRODUCTION

Measurements of stroke rates and stroke lengths are rapidly becoming common place in competitive swimming [1]. Reports from most major meets now routinely include computations of stroke rates and stroke lengths along with swimming speeds and split times for races [1]. One job of the coach is to help athletes find the optimum combination of rate components, such as lengths, speeds and times, that will allow athletes reproducing fast swims at each particular race distance. This task appears to be of fundamental importance in the training of elite swimmers.

Presently, for instance, the procedure used to determine the optimum combination of race components for each particular race distance is to ask elite swimmers to swim a series of 25 to 100 repeats at race speed using a variety of different stroke rates [1]. This simple technique, however, can be extremely laborious, expensive and time-consuming. Moreover, such analyses do not alter our understanding of elite swimmers' race patterns, through an inclusive analysis of the entire repertoire of race components, a continuing comprehensive window into the combination of race components that will result in the best performance of each event. What is required instead is a holistic approach to an analysis of elite swimmers' race patterns that focuses on illuminating order in the entire set of observations, allowing coaches to develop an integrated knowledge of the swimmers' race patterns being studied. We need to develop some fundamental tools to find the optimum combination of race components that will allow athletes to reproduce fast swims at a particular race distance.

[+] Ilene Chen: ilene@ee.usyd.edu.au TEL: +61 2 93517221 FAX: +61 2 93513847

In our previous works [2, 3], we have shown how cluster analysis can be used to monitor a number of elite swimmers’ race patterns at the same time. In the present work, we aim to show how large-scale cluster analysis can be applied to discover the best pacing strategy and the optimum combination of race components that will help athletes achieve their best performance at each particular race distance. To illustrate this approach, we perform a large-scale cluster analysis on a combined data of male and female finalists’ swimming speeds and stroke lengths, respectively, in the 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid. Several of the basic ideas for our approach are taken from our previous works [2, 3]. Our purpose in writing this dissertation has been to provide a comprehensible approach to discover the best pacing strategy and the optimum relationship among race components that will increase athletes’ chances to improve their times at each particular race distance. This technique as presented is referred to as retrospective cluster analysis of elite swimmers’ race patterns.

## 2. METHODS

**Source of data.** Pursuant to the data sharing policies, verification and reproduction policies, a total of eight data tables relating to male and female finalists’ race patterns in the 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid, were collected from a web site maintained by Professor Rein Haljand, Department of Kinesiology, Tallinn University. These data tables are available from Profession Rein Haljand’s web site at http://www.swim.ee/competition.

In this analysis, we have constructed combined data relating to male and female finalists’ race patterns on swimming speeds and stroke lengths, respectively, in the 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

**Hierarchical Clustering.** Details in the application of cluster analysis for studying underlying group structure on the basis of elite swimmers’ race patterns were given in our previous works [2, 3]. The similarity metric used here is the Euclidean distance, defined as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{d} \left( X_{ik} - X_{jk} \right)^2}$$

where $X_{ik}$ and $X_{jk}$ are the variable values for swimmers $i$ and $j$; the number of variables is $d$. In this analysis, the hierarchical clustering algorithm is used, which leads to a series of hierarchical groups, reported by tree-like dendrograms [4]. The similarity between two clusters is based on average linkage clustering, where the inter-cluster distances are defined as the average of the distance between all pairs of data points and members of a pair which are in distinct clusters [4].

**Displays.** The primary combined data relating to male and female finalists’ swimming speeds and stroke lengths, respectively, are reported graphically using heat maps with columns corresponding to swimmers and rows corresponding to observations on swimmers’ swimming speeds and stroke lengths, respectively. Here we use heat maps to plot the data relating to male and female swimmers’ swimming speeds and stroke lengths, respectively, on the 2-dimensional grids and use the rank-by-feature colour scale scheme to represent the magnitude of the data on the basis of their race patterns [5, 6]. In this analysis, medium data points are coloured black, data points with increasing ranks with reds of increasing intensity and decreasing ranks with greens of decreasing intensity. Then, outputs from average linkage hierarchical clustering are reported using tree-like dendrograms that are appended to the corresponding heat maps to represent the nature of the computed relationships among male and female finalists in the combined primary data using any ordering.

## 3. RESULTS

We apply average linkage hierarchical clustering to the combined data on male and female finalists’ swimming speeds (Figure 1) and stroke lengths (Figure 2), respectively, in the 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid. The order of the race record times commences from the fastest for the 200m freestyle finals across the 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid is: Hoogenband (01:44.9 in Berlin), Hoogenband (01:45.3 in Sydney), Thorpe (01:45.8 in Sydney), Davis (01:46.7 in Sydney), Rosolino (01:46.7 in Sydney), Brembilla (01:46.9 in Berlin), Rosolino (01:47.3 in Helsinki), Hoogenband (01:47.5 in Madrid), Hoogenband (01:47.6 in Helsinki), Palmer (01:47.8 in Sydney), Rosolino (01:48.0 in Berlin), Kapralov (01:48.3 in Madrid), Svoboda (01:48.4 in Berlin), Magnini (01:48.7 in Madrid), Salter (01:48.7 in Sydney), Rosolino (01:48.7 in Madrid), Barnier (01:48.8 in Berlin), Leveaux (01:48.8 in Madrid), Say (01:48.8 in Sydney), Svoboda (01:49.2 in Madrid), Wildeboer (01:49.2 in Berlin), Oikonomou (01:49.4 in Berlin), Palmer (01:49.5 in Helsinki), Hackett (01:49.5 in Sydney), Herbst (01:49.6 in Berlin), Svoboda (01:50.2 in Helsinki), Herbst (01:50.4 in Helsinki), Wildeboer (01:50.9 in Madrid), Kapralov (01:51.4 in Helsinki), Carstensen (01:51.4 in Helsinki), Carstensen (01:52.2 in Madrid), and Arnarson (01:52.3 in Helsinki), Vanalmsick (01:56.6 in Berlin), Potec (01:57.8 in Berlin), Popchanka (01:57.9 in Berlin), Potec (01:58.2 in Madrid), O'neill (01:58.24 in Sydney), Figues (01:58.3 in Madrid), Moravcova (01:58.32 in Sydney), Poll (01:58.81 in Sydney), Tchemezova (01:58.86 in Sydney), Kielglass (01:58.86 in Sydney), Baranovskaya (01:59.3 in Sydney), Potec (01:59.46 in Sydney), Nabanouskaya (01:59.5 in Helsinki), Lillhage (01:59.5 in Madrid), Wang (01:59.55 in Sydney), Figues (01:59.6 in Berlin), Moravcova (02:00.1 in Helsinki), Potec (02:00.3 in Helsinki), Pelleggrin (02:00.3 in Madrid), Rouba (02:00.5 in Madrid), Veldhuis (02:00.6 in Madrid), Figues (02:00.7 in Helsinki), Popchanka (02:01.2 in Madrid), Roca (02:01.3 in Berlin), Rouba (02:01.5 in Berlin), Hjorthhansen (02:01.6 in Berlin), Roca (02: 01.7 in Helsinki), Koechoven (02:01.74 in Helsinki), Caballero (02:02 in Madrid), Parise (02:02.2 in Helsinki), Vanrooijen (02:02.2 in Berlin), Vlieghuis (02:02.7 in Helsinki).

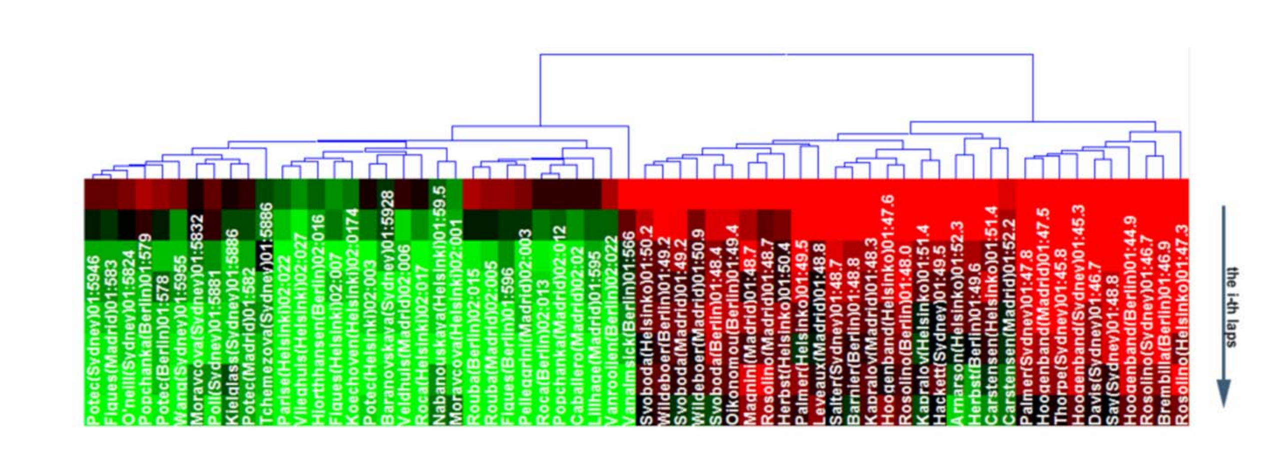

Figure 1 Clustered display of data from the male and female finalists' swimming speeds for the 200m freestyle finals across the 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

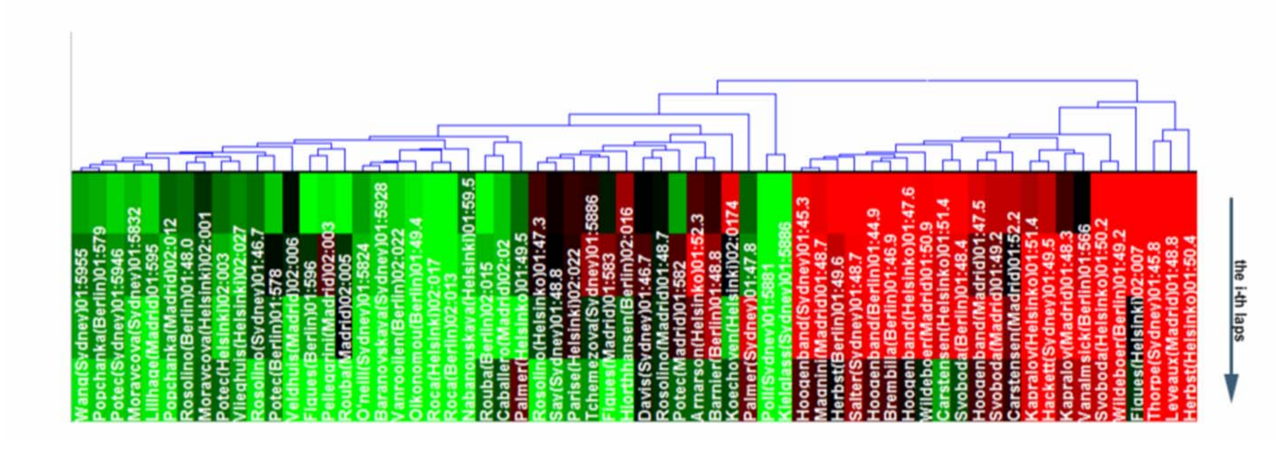

Figure 2 Clustered display of data from the male and female finalists' stroke lengths for the 200m freestyle finals across the 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid.

Arranging male and female finalists' swimming speeds and stroke lengths, respectively, over their entire race record times through each of the i-th laps across several major meets allows us to recognise the race patterns of interest, readily zoom in on the detailed race patterns, and characterise the swimmers contributing to those race patterns. What we have found repeatedly is that swimmers represent more than one heat map next to which the grid elements are clustered or in the immediate vicinity of each other, indicating that swimmers who share similar race patterns tend to be clustered together [3]. On the other hand, individual race patterns can also be distinguished from all others because of the considerable dissimilarities in their race patterns [3]. Therefore, the visual representation of elite swimmers' race patterns using such images mimics the observed race patterns.

We consider the two-group arrangements given by average linkage hierarchical clustering corresponding to male and female finalists' swimming speeds and stroke lengths, respectively, in the 200m freestyle finals across 2000 European Championships in Helsinki, Sydney 2000 Summer Olympics, 2002 European Championships in Berlin, and 2004 European Championships in Madrid. Corresponding to male and female finalists' swimming speeds, the two-group arrangement given by average linkage hierarchical clustering is:

Cluster 1: all of the male finalists;

Cluster 2: all of the female finalists.

Corresponding to male and female finalists' stroke lengths, the two-group arrangement given by average linkage hierarchical clustering is:

Cluster 1: Thorpe (01:45.8 in Sydney), Leveaux (01:48.8 in Madrid), and Herbst (01:50.4 in Helsinki), Hoogenband (01:47.6 in Helsinki), Hoogenband (01:45.3 in Sydney), Hoogenband (01:44.9 in Berlin), Carstensen (01:51.4 in Helsinki), Carstensen (01:52.2 in Madrid), Magnini (01:48.7 in Madrid), Herbst (01:49.6 in Sydney), Salter (01:48.7 in Sydney), Brembilla (01:46.9 in Berlin), Wildeboer (01:50.9 in Madrid), Svoboda (01:49.2 in Madrid), Kapralov (01:51.4 in Helsinki), Vanalmsick (01:56.6 in Berlin), Svoboda (01:50.2 in Helsinki), Wildeboer (01:49.2 in Berlin), Figues (02:00.7 in Helsinki).

Cluster 2: Male and female finalists other than those in Cluster 1.

When we consider the two-group arrangement corresponding to the finalists' swimming speeds, the resultant tree-like dendrogram exhibits that finalists' of the same gender tend to be clustered together. On the other hand, when we consider the two-group arrangement corresponding to the finalists' stroke lengths, the resultant tree-like dendrogram shows that finalists of different genders are clustered together. Consequently, male and female swimmers may use similar race patterns on stroke lengths in their race although the swimming speeds of the males are generally somewhat faster.

Moreover, we can explore relationships among swimmers on the basis of their race patterns by choosing different cut-off similarity points given by the average linkage hierarchical clustering [2, 3]. If we consider, for example, the four-group arrangement corresponding to the finalists' stroke lengths, the four-group solution is:

Cluster 1: Thorpe (01:45.8 in Sydney), Leveaux (01:488 in Madrid), and Herbst (01:50.4 in Helsinki);

Cluster 2: Hoogenband (01:47.6 in Helsinki), Hoogenband (01:45.3 in Sydney), Hoogenband (01:44.9 in Berlin), Carstensen (01:51.4 in Helsinki), Carstensen (01:52.2 in Madrid), Magnini (01:48.7 in Madrid), Herbst (01:49.6 in Sydney), Salter (01:48.7 in Sydney), Brembilla (01:46.9 in Berlin), Wildeboer (01:50.9 in Madrid), Svoboda (01:49.2 in Madrid), Kapralov (01:51.4 in Helsinki), Vanalmsick (01:56.6 in Berlin), Svoboda (01:50.2 in Helsinki), Wildeboer (01:49.2 in Berlin), Figues (02:00.7 in Helsinki);

Cluster 3: Poll (01:58.81 in Sydney), Kielglass (01:58.86 in Sydney);

Cluster 4: Swimmers other than those in Clusters 1, 2 and 3.

The groups corresponding to the finalists' race patterns on stroke lengths show that Thorpe's stroke lengths when he swam freestyle 200m with a time of 01:45.8 minutes at Sydney 2000 Summer Olympics, Herbst's stroke lengths when he swam freestyle 200m with a time of 01:50.4 minutes at 2000 European Championships in Helsinki, and Leveaux's stroke lengths when he swam freestyle 200m with a time of 01:48.8 minutes at 2004 European Championships in Madrid are arranged into a single cluster containing all of them. This means that their race patterns on stroke lengths over their entire race record times through each of the i-th laps are distinguishing and have particular similarities. In other words, Herbst and Leveaux have developed favorable patterns on stroke lengths, as excellent as Thorpe's, over their entire race record times through each of the i-th laps. On the basis of our present observations, it is likely that Herbst and Leveaux can reproduce fast swims by increasing their stroke rates and simultaneously maintaining their stroke lengths at or near previous levels in the 200m freestyle race distance events.

We can also use large-scale cluster analysis to find the best pacing strategy that would assist athletes to improve their times at each particular race distance. If we consider, for example, the three-group arrangement corresponding to the finalists' pace, the three-group solution is:

Cluster 1: Rosolino (01:47.3 in Helsinki), Brembilla (01:46.9 in Berlin), Rosolino (01:46.7 in Sydney), Hoogenband (01:44.9 in Berlin), Say (01:48.8 in Sydney), Davis (01:46.7 in Sydney), Hoogenband (01:45.3 in Sydney), Thorpe (01:45.8 in Sydney), Hoogenband (01:47.5 in Madrid), and Palmer (01:47.8 in Sydney).

Cluster 2: Male finalists other than those in Cluster 1;

Cluster 3: All female finalists.

Since Rosolino (01:47.3 in Helsinki), Brembilla (01:46.9 in Berlin), Rosolino (01:46.7 in Sydney), Hoogenband (01:44.9 in Berlin), Say (01:48.8 in Sydney), Davis (01:46.7 in Sydney), Hoogenband (01:45.3 in Sydney), Thorpe (01:45.8 in Sydney), Hoogenband (01:47.5 in Madrid), and Palmer (01:47.8 in Sydney) are arranged into an individual cluster, this means that their paces have particular similarity. Among all of them, Hoogenband's race results in the 2002 European Championships in Berlin are superior to the other swimmers' events. Therefore, Hoogenband's pacing when he swam freestyle 200m with a time of 01:44.9 minutes at European Championships in Berlin is identified to be the best pacing strategy in the 200m freestyle race distance events. It is, therefore, possible to enable Palmer, Thorpe, Davis, Say, Brembilla and Rosolino to improve their times by adjusting their current paces at or near the levels of Hoogenhand's paces at the 2002 European Championships in Berlin.

## 4. DISCUSSION

We have shown how large-scale cluster analysis can be applied to discover the best pacing strategy and the optimum combination of race components that will help athletes achieve their best performance at each particular race distance in a natural way. The analysis of the data described in the current work has a drastically diverse and expanded scope. We outline three main findings: Firstly, based on the resultant tree-like dendrograms from cluster analysis, we find that male and female swimmers may use similar patterns on stroke lengths in their race whereas the swimming speeds of males are generally somewhat faster. Secondly, we demonstrate a potential application of large-scale cluster analysis to find the optimum relationship between stroke length and swimming speed that will allow German national swimmer, Stefan Herbst, and French national swimmer, Amaury Leveaux, to reproduce fast swims in the 200m freestyle race distance events. Thirdly, we demonstrate a potential application of large-scale cluster analysis to find the best pacing strategy that will allow British national swimmer, Paul Palmer, Australian national swimmer, Ian Thorpe, American national swimmer, Joshua Davis, Canadian national swimmer, Rick Say, Italian national swimmer, Emiliano Brembilla and Massimiliano Rosolino, to improve their times in the 200m freestyle race distance events.

What we have found to be the most valuable feature of the approach described here is that it allows coaches and athletes to discover the best pacing strategy and the optimum combination of race components that will help athletes achieve their best performance at each particular race distance. Unlike Competition Analysis [7] or Race Analysis [8, 9] in current use, the results from large-scale cluster analysis are displayed using tree-like dendrograms that are appended to the corresponding heat maps, rather than numbers, which are more insightful to coaches and athletes in finding the best pacing strategy and the optimum relationships among race components that will largely increase athletes' chance of improving their times. These visual displays present all the quantitative information, but convey the information to our human brains more efficiently by means of a high-bandwidth channel than a "number-reading" channel [10]. Since our human brains are not well adapted to assimilate quantitative data by reading digits, it would meet with a great deal of enthusiasm from coaches if they can find this information from such tree-like dendrograms [10].

Finally, we have seen that the success of retrospective cluster analysis has given us confidence to discover the best pacing strategy and optimum combinations of race components that will help athletes achieve their best performance at each particular race distance in a natural way.

## 5. ACKNOWLEDGMENT

This work is in part supported by a grant from City University of Hong Kong (Project CityU 9610034).

## 6. REFERENCES

[1] E. W. Maglischo, Stroke rates and stroke lengths, *Swimming Fastest*. Human Kinetics, 605-701, 2003

[2] I. Chen, H. Homma, C. Jin, H. Yan. Identifcation of elite wimmers’ race patterns using cluster analysis. *Journal of Sport Science and Coaching*, 203-302, 2007

[3] I. Chen, H. Homma, C. Jin, H. Yan. Clustering and display of elite swimmers’ race patterns across various comparable criteria at the same time. *International Journal of Sport Science and Engineering*, 129-135, 2007

[4] B. S. Everitt, Cluster analysis: classifying countries in terms of the athletic prowess of their women, *Statistical Analyses using S-plus*, 104-115, 1994

[5] Seo, J. and B. Shneideman. A Rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projection. *Proceeding to IEEE Information Visualization*, 65-72, 2004

[6] Hierarchical Clustering Explorer 3.0, http://www.cs.umd.edu/hcil/hce/hce3.html.

[7] R. Haljand. Competition analysis in swimming. *The World of Swimming*, February, 1993.

[8] Cossor, J. and Mason, B., Swim Start Performance at the Sydney 2000 Summer Olympic. http://www.coachesinfo.com/category/swimming/143.

[9] Cossor, J. and Mason, B., Swim Turn Performance at the Sydney 2000 Summer Olympic. http://coachsinfo.com/category/swimming/144.

[10] Eisen, M. B., Spellman, P. T., Brown, P. O. and Potstein, D., *Cluster Analysis and Display of Genome-Wide Expression Patterns*. *Proceeding to National Academic Science, USA*, 14863-14868, 1998.