Sports analytics for professional speed skating

Knobbe, A.J.; Orie, J.; Hofman, N.; Burgh, B. van der; De Gouveia da Costa Cachucho, R.E.

Data Min Knowl Disc (2017) 31:1872–1902
DOI 10.1007/s10618-017-0512-3

Sports analytics for professional speed skating

Arno Knobbe[1,2] · **Jac Orie**[3,4] · **Nico Hofman**[3,4,5] · **Benjamin van der Burgh**[1,6] · **Ricardo Cachucho**[1]

Received: 7 March 2016 / Accepted: 12 May 2017 / Published online: 27 May 2017

Abstract In elite sports, training schedules are becoming increasingly complex, and a large number of parameters of such schedules need to be tuned to the specific physique of a given athlete. In this paper, we describe how extensive analysis of historical data can help optimise these parameters, and how possible pitfalls of under- and overtraining in the past can be avoided in future schedules. We treat the series of exercises an athlete undergoes as a discrete sequence of attributed events, that can be aggregated in various ways, to capture the many ways in which an athlete can prepare for an important test event. We report on a cooperation with the elite speed skating team LottoNL-Jumbo, who have recorded detailed training data over the last 15 years. The aim of the project was to analyse this potential source of knowledge, and extract actionable and interpretable patterns that can provide input to future improvements in training. We present two alternative techniques to aggregate sequences of exercises into a combined, long-term training effect, one of which based on a sliding window, and one based on a physiological model of how the body responds to exercise. Next, we use both linear modelling and Subgroup Discovery to extract meaningful models of the data.

Keywords Sports analytics · Speed skating · Physiological modelling · Subgroup discovery · Sequence mining

✉ Arno Knobbe
a.j.knobbe@liacs.leidenuniv.nl

1 LIACS, Universiteit Leiden, Leiden, The Netherlands

2 Hogeschool van Amsterdam, Amsterdam, The Netherlands

3 LottoNL-Jumbo, Amsterdam, The Netherlands

4 Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

5 Paramedisch Adviescentrum Aalsmeer, Aalsmeer, The Netherlands

6 Kiminkii, Houten, The Netherlands

1 Introduction

This paper describes research challenges related to a recent Sports Analytics project between a leading Dutch professional speed skating team and data scientists from Universiteit Leiden and Hogeschool van Amsterdam. During its history, the athletic team, currently called LottoNL-Jumbo,[1] has included numerous Dutch skaters that competed at the European, world, and Olympic level, and presently includes a world record holder and two Olympic medalists. The project involved 15 years of detailed training data kept by the coach of the team (second author of this paper) with the aim of improving the training program and further optimising the performance of current and future skaters of the team. In this paper, we report on the data science techniques required to analyse this non-trivial data, and showcase findings for specific athletes. A number of novel techniques are introduced to deal with the specifics of the recorded data, and to produce interpretable and actionable results that can help fine-tune the training programs.

Speed skating is a winter sport where athletes compete on skates to cover a given distance on an oval (indoor) ice rink. Although there are various disciplines, including marathons and short-track speed skating, we focus here on long-track speed skating, which involves a 400 m oval track with two lanes. Events over multiple distances exist, ranging from 500 to 10,000 m (for men), with each skater typically specialising in one or two distances, depending on their physiology and training. Although each race in an event involves two skaters, the final standing is determined by the overall ranking of times of all participants. This effectively makes each race a time trial, where the outcome of a given skater is only determined by their own performance (ignoring for the moment the interference due to having to switch lanes). From a data mining point of view, this is attractive since all results can be assumed independent, and one can simply collect all race results of an athlete without having to consider the influence of the 'opponent'.

A typical season for a professional speed skater lasts eleven months, with one month of break at the end of the (winter) season. Of these eleven months, roughly five months involve the actual competition season, such that six months of the season are spent in relative isolation, in preparation for competition. Training consists of a complex mixture of training impulses of various types, with an emphasis on physical rather than technical training (in part caused by the lack of good quality ice during the summer). A training program is intentionally mixed, with different types of activities and periods of roughly 2 weeks in which certain physical loads are emphasised or avoided. The aim here is to trigger different physiological mechanisms and to avoid exhausting or desensitising certain parts of the system by a too monotonous training regimen. Aside from training, a skater's routine consists of regular test moments, which during winter are simply the races, and during the preparation season standardised physical tests of various types (Åstrand and Ryhming 1954; Vandewalle et al. 1987).

The available data, painstakingly collected by the coach, involves primarily descriptions of the daily training activities, partly structured and partly free text. The structured

[1] Commercial speed skating teams need to regularly change their name, to reflect the sponsors. Over the last 2 years, the team has been sponsored by the Dutch Lotto and supermarket chain Jumbo.

part of the description is very consistent, and captures a classification of the nature of the training (e.g. "*cycling extensive endurance*"), as well as numeric values indicating the duration (in minutes) and intensity (on a subjective scale of 1–10) of the session. Training data is specific to individual athletes, and the intensity of the training was obtained from the athlete, post hoc. With six training days per week, and potentially two sessions per day, this amounts to roughly 450 sessions per season, making for a substantial data collection per athlete/season. Next, several datasets are available that capture test events. These include race results that were scraped from the internet,[2] as well as physical test results, which were primarily performed during summer to get an indication of the level of fitness and the efficacy of the training program thus far. Note that several different datasets and associated mining tasks can be derived from this combined data, depending on the choice of entity involved. For example, one could produce a dataset describing races of a specific distance, for a given skater, with the race outcome as dependent variable, and the training activities prior to each race as independent variables. Equally useful, one could take all outcomes of a specific test as entities, have a combination of race and test results, or even combine different distances or athletes. In our experiments, we will consider several of these possibilities. However, in each case the problem is essentially a *regression task*, since the target variable is numeric [e.g. time on the 500 m or the VO_2max value (Kenney et al. 2015; McArdle et al. 2014) established with a cycling test].

While evidently being in the regression domain, it is not immediately clear what the independent variables of our task might be. Clearly, the independent variables should capture aspects of the training program prior to the events in question. However, the available data is a long sequence of events with a small number of characteristics (type, duration, intensity, …), so some form of transformation is necessary to arrive at an attribute-value representation that is amenable to main-stream regression analysis. In this paper, we take an *aggregation-based feature construction* approach, inspired by earlier work in Cachucho et al. (2014), in order to derive a fairly extensive set of features that capture the preparation (training, but also absence thereof) from various angles, for example focussing on specific periods prior to the test event, or on specific intensity zones. In its basic form, the aggregation will take place over windows of varying lengths (ranging from one day to several weeks) using different aggregate functions and variables, with specifiers such as training type and intensity zones. In a more elaborate approach, developed for this specific purpose, the aggregation will take the form of convolution with a physiology-inspired kernel consisting of several exponential decay functions of varying half times. This kernel is inspired by the so-called *Fitness-Fatigue model* (Calvert et al. 1976), that tries to capture how the human body responds to a specific training impulse over the course of time.

After having obtained a suitable attribute-value representation with potentially predictive features, the next challenge is to produce meaningful models from this dataset. The overall aim of this project is to provide the coaching staff with easy-to-understand, actionable pointers as to how to fine-tune the training routines, and avoid pitfalls of under and over-training. Therefore, we specifically intend to discover interpretable

[2] http://www.osta.nl.

patterns, that are relatively easy to understand by the domain experts, and ideally do not involve a great many variables. We will be working with two types of regression techniques. Assuming mostly linear dependencies between the aggregated features and the target variable, *regularised linear regression* methods such as LASSO (Friedman et al. 2010; Tibshirani 1994) are attractive since they select features and produce relatively concise models of the data. However, with the physiological domain at hand, it is likely that non-linear dependencies will also exist, and rather, one expects thresholds to exist on the features, where too large or too low a value (e.g. training load) will produce sub-optimal results. For such phenomena, we expect *Subgroup Discovery* (Klösgen 2002; Grosskreutz and Rüping 2009; Atzmüller and Lemmerich 2009; Pieters et al. 2010) to produce more useful results.

As a by-product of this project, we have developed ways of connecting the classical linear regression approaches to the regression setting of Subgroup Discovery, thus making linear models and subgroups comparable. We will show how subgroups can be interpreted as (potentially higher-dimensional) step functions such that their fit can be compared with that of the linear function. Possibly, this may require the linear models to be of low dimension, even involving only single variables, but this is actually attractive from an interpretation point of view, as argued above. Treating linear models and subgroups similarly also allows us to apply statistical tests to either, and compare the outcome of both approaches on significance.

This Sports Analytics paper has two sides. On the Sports side, we present some interesting findings that are of practical relevance to the team, with the following key contributions:

- The application of the Fitness-Fatigue model and the fitting of this model to individual skaters, where the parameters of the model convey key properties of each skater.
- Various demonstrative, interpretable patterns concerning improved training practices.
- The presentation of results relating to (1) competitions, and (2) physical tests.
- The capability to produce detailed findings for other skaters.

On the Analytics side, we introduce a number of new ways to exploit detailed training data, of relevance not just in the speed skating discipline, and in some cases applicable to other analytics domains also. Key contributions are:

- Introduction of (conditional) aggregation as a way of aggregating discrete sequences of events, and producing a range of features that capture various aspects of those sequences.
- Aggregation by means of two options: one that is easy to compute and interpret (uniform window), one that is more physiologically plausible, and at the same time harder to compute (the Fitness-Fatigue model).
- Application of linear modelling and Subgroup Discovery in order to select key features and produce interpretable models.
- Introduction of a novel SD quality measure, Explained Variance.
- Evaluation of models in terms of R^2 and p values, that makes linear models and subgroups immediately comparable.

2 Speed skating and sports analytics

The (long-track) speed skating takes place on an oval track 400 m in length. Races are typically held with two participants at each time (skating in separate lanes), but each participant is ranked on their individual time. Both men and women compete, in separate competitions. Races come in various distances, but the most common distances at major events are 500, 1000, 1500, 3000, 5000 and 10,000 m, of which the last distance only applies to men, who in turn do not skate the 3000 m. These disciplines are usually divided into sprint, medium, and long distance, and skaters typically specialise in one of these, and compete in only a few of these disciplines, although participation is facultative. Each category requires a specific type of physiology, which explains the specialisation of athletes. Furthermore, each distance requires a specific type of training, and exercises for one distance may actually harm the performance on other distances (McArdle et al. 2014; Kenney et al. 2015). This implies that our analysis will often be specific to a small number of similar distances, or even be specific to individual athletes. Since the training programs are well-developed, and the senior athletes will have several years of experience working with the coach, the produced findings may be subtle, which will often call for an athlete-specific approach.

Even though races for a specific event can be considered time trials, there will be a level of variance in the race results that cannot be explained by differences in race preparation and training. It is a well-know fact within speed skating that times are determined to a reasonable degree by the ice rink. First of all, the ice properties will differ from one venue to the next, and top venues (where major events are organised) will have better ice maintenance techniques and experience. Another factor that influences the times, besides ice quality, is the altitude of the venue, with higher ice rinks tending to be faster due to the lower air resistance. The top four of fastest venues consists of Salt Lake City, USA (1425 m elevation), Calgary, Canada (1034 m), Heerenveen, the Netherlands (0 m), and Sochi, Russia (5 m). In order to compensate for the difference in rink speeds, we have opted to work with *relative times* rather than recorded times.

Definition 1 (*Relative time*) For a race of distance d, by a skater of gender g, at ice rink r, finishing in time t (in seconds), the *relative time* t_{rel} is defined as

$$t_{rel} = t / t_{rec}(d, g, r)$$

where $t_{rec}(d, g, r)$ is the record for a specific discipline and ice rink.

Relative time will produce race results slightly above 1.0 or exactly 1.0 if a race either produced or replicated a rink record. The use of relative time not only allows comparisons between results at different venues, but theoretically also comparisons between results in different disciplines or even between men and women, although one might be comparing apples and oranges here on other aspects of the data. The rink records were scraped from the Dutch Wikipedia page[3] that collects local records. Note that the use of records to estimate the speed of a rink is not flawless. First of all, records are continuously subject to improvement, such that the definition of relative time is

[3] https://nl.wikipedia.org/wiki/Lijst_van_snelste_ijsbanen_ter_wereld.

sometimes problematic. Next, some venues rarely host international events, such that their records do not fairly represent the theoretical speed of the rink, and actually produce under-estimates of the relative times athletes set (meaning they appear faster than they are). In order to avoid constantly having to update the list of records and subsequent analyses, which is rather time-consuming, we choose to fix the records at a particular point in time. A minor side effect of this is that some newer results may actually have a relative time below 1.0. Our collected list of international rink records will be made available[4] as part of this publication.

3 Feature construction by aggregation

As explained in the introduction, the training data takes the form of a sequence of annotated events, corresponding to the individual exercises an athlete performs. While being valuable information, this sequential representation will require certain transformations in order to elicit general characteristics of an athlete's preparation for a race. Individual exercises will generally not play a deciding role in the outcome (unless of very extreme nature), and it is the combined nature of exercises that determines the effect of the training program. Therefore, some form of aggregation is required to draw out the various aspects of training that potentially play a role. Although there is a large body of knowledge about the effect of certain types of exercise in the sports physiology literature, it is still uncertain what aspects of training and preparation determine the variance in race results that still remains, as for example exemplified in Fig. 1. For this reason, our feature construction approach will include a rather large collection of features, with the aim of including many angles and leaving room for discovery in later stages of the analysis. Furthermore, it is not quite clear how long the effect of specific exercises lasts for individual athletes, and thus what period prior to the test event should actually be included in the aggregation. Our set of constructed features will therefore involve time windows of various lengths, ranging from one day to several weeks.

In this paper, we will consider two general aggregation approaches, the first of which involves uniform aggregation over the various windows.

3.1 Uniform aggregation

Before defining the notion of a (time) window, we first formalise the events to be aggregated, as they appear in the data of our application.

Definition 2 (*Exercise*) An *exercise* is defined as a tuple $e = (t, ampm, dur, int, load)$, where t is the date of the exercise, $ampm$ is a binary variable indicating the morning or afternoon session. Numeric values dur, int, and $load$, indicate the duration (in minutes), the subjective intensity (on a scale of 1–10), and the load (in intensity-minutes) of the exercise, respectively.

The three crucial numeric attributes of an exercise specify the following:

[4] http://datamining.liacs.nl/rink-records.txt.

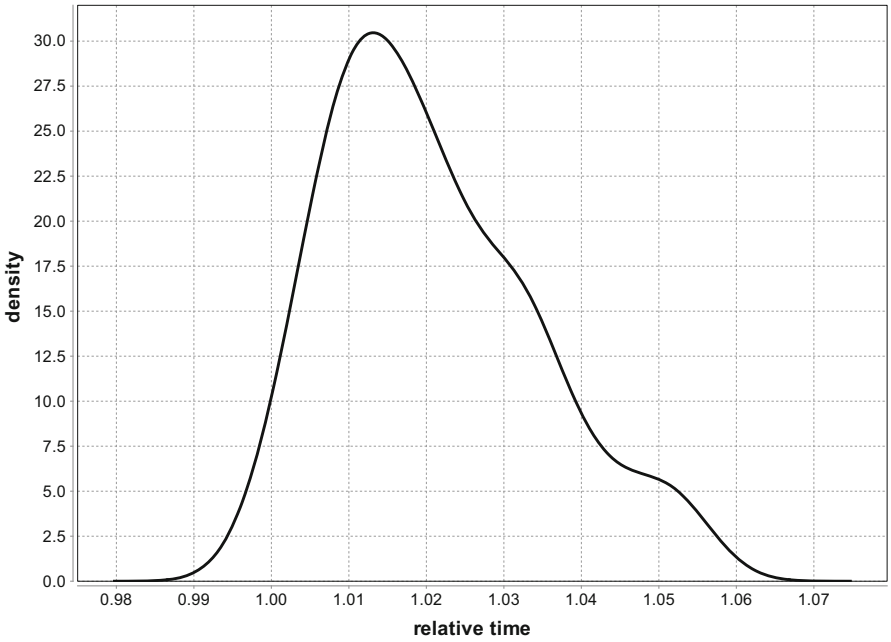

Fig. 1 Distribution of relative times over 1000 m as estimated by kernel density estimation (based on 75 results, with a Gaussian kernel of $\sigma = 0.0051$). The probability density estimate continuing below 1.0 is an artefact of the KDE. The best time in the list is actually a rink record in the Hague (the Netherlands) at 1.0

- The *duration* simply specifies the length of the exercise. Durations tend to be rounded to quarters of hours (especially for longer exercises), but this is deliberate, and athletes generally adhere to the required duration.
- The *intensity* indicates how heavy the exercise was, as perceived by the athlete, with 10 being the intensity of a race itself. During training, values of 9 or 10 are rare. Although intensity is a subjective measure, the athletes are very used to it, and will rate specific trainings fairly consistently.
- The *load* is defined as the product of duration and intensity, with the intention of capturing the total energy expenditure during the exercise. Although load is actually a derived attribute (it does not appear in our normalised database, for that reason), we include it in the definition of an exercise because it plays a crucial role both in the modelling as in the reasoning behind the training program.[5]

Note that the races themselves also appear as exercises in the database, since it is crucial to include the training load produced by such intense events, when considering the preparation for other races. In speed skating, several races often take place in a single weekend, such that later races are influenced by earlier ones.

[5] Note that the definition in terms of a product of duration and intensity might be too simplistic, since neither duration nor intensity may be a linear scale. Doubling the length of an exercise may have an exaggerated effect if the intensity is (too) high, and doubling the intensity makes the exercise entirely different in nature, addressing different metabolic systems.

Definition 3 (*Time Window*) A *(time) window* $w_{t,m}$ is a set of exercises $e_1, \ldots, e_n$, such that all dates $e_i.t$ are before t, and not more than $2m-1$ days before t: $t-2m+1 \leq e_i.t < t$.

Note that day t itself is not part of the window. For reasons that will become clear in later sections, we have opted to define the length of a window in terms of its middle m, essentially indicating the 'centre of mass' of the window. A window $w_{t,1}$ will thus include the 1 day prior to t, $w_{t,2}$ indicates the 3 days prior to t, and so on.

For a window w, the following primitive aggregates will be considered:

Count	Simply the number of exercises in w: $\|w\|$.
Sum(duration)	The sum of durations of the exercises in w: $\sum_i e_i.dur$.
Sum(intensity)	The sum of intensities of the exercises in w: $\sum_i e_i.int$.
Sum(load)	The sum of loads of the exercises in w: $\sum_i e_i.load$.
Avg(duration)	The average duration of the exercises in w: $\sum_i e_i.dur/\|w\|$.
Avg(intensity)	The average intensity of the exercises in w: $\sum_i e_i.int/\|w\|$.
Avg(load)	The average load of the exercises in w: $\sum_i e_i.load/\|w\|$.
Max(duration)	The maximum duration of the exercises in w: $\max_i e_i.dur$.
Max(intensity)	The maximum intensity of the exercises in w: $\max_i e_i.int$.
Max(load)	The maximum load of the exercises in w: $\max_i e_i.load$.

Aggregation using the minimum was deemed senseless, since a very light training has little effect, and one could interpret daily rest periods as very light exercises anyway.

3.1.1 Specifiers

Each primitive aggregate listed above can be applied to all the exercises in a given window, or just to subsets of exercises from specific categories. These subsets are specified by what we will refer to as specifiers. We apply the following specifiers:

Morning/afternoon sessions By specifying *am*, *pm* or no specifier at all, the aggregate can include only the morning sessions, only the afternoon sessions, or all sessions, respectively. Note that during the winter, the coach will plan exercises on the ice in the morning, and alternative training in the afternoon, so distinguishing between those may be fruitful.

Intensity intervals Exercises at different intensities will trigger different parts of the system, and hence will produce a different training stimulus. As specifier, we optionally select only exercises within specific intensity intervals $[l, u]$, where $l \in [1, 10]$ and $u \in [l, 10]$.

Note that each type of specifier will introduce multiple variants of the primitive aggregates. For *ampm*, adding specifiers will raise the number of aggregates (per window size) from 10 to 30. For the intensity-intervals, there will be $1/2 \cdot 10 \cdot (10+1) = 55$ variants of each primitive. However they are only applied to the 4 primitive aggregates that do not involve intensity and load, producing $55 \cdot 4 = 220$ aggregates. In order to avoid combinatorial explosion of the aggregate collection, we do not include combinations of specifiers (such as low intensities in morning sessions). In total, there will be 250 aggregates per window.

In "Appendix A", details are provided about how the required form of aggregation can be achieved in a relational database using SQL statements. This appendix explains how the data is modelled as two tables, one listing all results of races (`Result`), and one listing all exercises (`Exercise`). There is an implicit one-to-many relationship between these two tables, and joining and subsequent aggregation (by means of a `GROUP BY` statement) will combine details of individual exercises into several aggregated features on the level of `Result`.

3.1.2 Aggregation and convolution

Observe that such uniform aggregation over a window can be seen as a form of convolution with a rectangular kernel (Stranneby and Walker 2004). The convolution of a time series $x(t)$ (in this case any of the training parameters that are aggregated) applies a kernel to the series to obtain a new series $y(t)$ as follows:

$$y(t) = h * x(t) = \sum_{i=-\infty}^{\infty} h(i)x(t-i)$$

Here, h refers to the kernel, which is required to sum to 1 over its domain. In the case of a uniform window, the kernel is essentially a rectangular function (remember that $2m-1$ is the length of the window):

$$h(t) = \begin{cases} 1/(2m-1) & \text{if} \quad 0 \leq t \leq 2m-1 \\ 0 & \text{otherwise} \end{cases}$$

Since the rectangular kernel is zero over a large part of its domain, the convoluted function $y(t)$ can be computed much faster. In the next section, we will introduce a kernel that is both more natural and more expensive to compute.

3.2 The Fitness-Fatigue model

Although uniform aggregation is intuitive and straightforward to implement (and as we will see, provides fairly good models), it is somewhat unnatural. First of all, it is unlikely that all exercises over a period of, say, 4 weeks will have the same influence on the level of fitness at a race. Rather, one would expect that exercises several weeks ago have a much smaller influence than more recent ones. Second, the hard distinction between an exercise 28 days ago, and one 29 days ago seems unnatural, and may introduce minor artefacts in the constructed features. Finally, there is a general pattern where the initial effect of an exercise is negative, while after a short period of rest and recuperation, the effect is positive. Ideally, the aggregated features should exhibit such behaviour.

In this section, we introduce a type of aggregation based on convolution with a more natural gradually progressing kernel. We will use a multi-component kernel that is taken from the physiology literature (Calvert et al. 1976) and aims to model the

complex way in which a human body responds to exercise by initial fatigue, followed by a slight improvement in performance, the effects of which die out gradually over the course of several week, returning to a state of fitness comparable to that prior to the exercise.

The core of this kernel is based on the *exponential decay* function, as follows:

$$h_e(m) = e^{-\lambda m}, \quad m \geq 0$$

The parameter λ here determines the speed with which this kernel decays towards zero, in other words, the speed with which the effects of exercise diminish over time. Although the exponential decay is defined in terms of λ (with unit s^{-1}), we will primarily define a specific kernel in terms of the parameter τ (in units s, or more conveniently, in days), which corresponds to the 'mean lifetime' of the kernel, and as such can be interpreted as the centre of mass of the kernel. This makes this parameter immediately comparable to parameter m of a uniform window, which is also the centre of mass of the kernel. The simple relationship between τ and λ is as follows:

$$\tau = 1/\lambda$$

The exponential decay function effectively models the diminishing positive effect of an exercise as time passes. However, it does not include the tiring effects of exercise in the few days after training, which may outweigh the positive influence of training. For this reason, Calvert et al. (1976) introduced the so-called *Fitness-Fatigue* model, which models these two effects as two components of a kernel with different weights and different decay factors, as follows:

$$h_2(m) = e^{-\lambda_{fit} m} - K e^{-\lambda_{fat} m}, \quad m \geq 0$$

λ_{fit} determines the speed with which the positive effects of training (the *fitness*) diminish, and typically corresponds to a τ_{fit} in the order of months, while λ_{fat} determines the shape of the fatigue curve immediately after the exercise. The associated τ_{fat} is typically in the range of two weeks. Initially, the influence of fatigue is about twice as big as that of the improved fitness (determined by the value of K).

The fitness in the above two-component model is assumed to be immediately improved after the exercise, which in practice is not the case. The desired adaptation in various metabolic systems will not take effect until several days after the exercise, such that the fitness will need to be modeled with an additional component (Calvert et al. 1976), producing the following three-component kernel:

$$h_{ff}(m) = (e^{-\lambda_{fit} m} - e^{-\lambda_{del} m}) - K e^{-\lambda_{fat} m}, \quad m \geq 0$$

where λ_{del} affects the exponential function that reduces the initial fitness. In Fig. 2, the combination of fitness and fatigue into this kernel is demonstrated. In Calvert et al. (1976), values are given for the associated parameters, obtained by fitting the convolved function to athletic data, producing the values below. Although these values seem reasonable, they will be athlete- and specialism-specific, such that we will fit

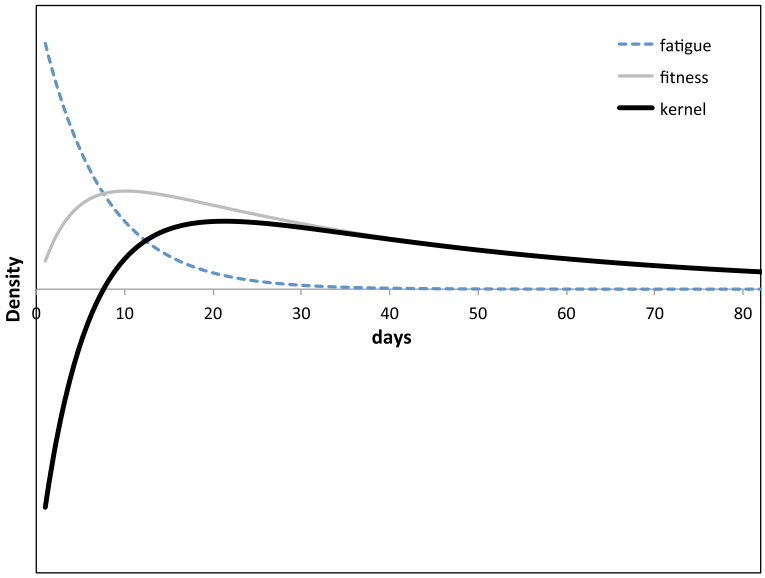

Fig. 2 The three-component Fitness-Fatigue kernel (in *solid black*) as a function of time after the exercise (in days). The *fitness* and *fatigue* parts are also shown, in *solid grey* and *dashed blue*, respectively (Color figure online)

these values to specific datasets collected, in the experimental section. The published values for the parameters are as follows:

$$\tau_{fit} = 50 \text{ days}, \quad \tau_{del} = 5 \text{ days}, \quad \tau_{fat} = 15 \text{ days}, \quad K = 2.0$$

"Appendix B" presents implementation details of the Fitness-Fatigue model in SQL.

4 Modelling approaches

4.1 Regularised linear regression

In the previous section, we explained the procedures to build large sets of interpretable features about the training, that might be able to explain the target variables of performance. These target variables might be a linear function of a subset of the aggregate features, but we do not know which ones beforehand. In order to find a good subset of aggregate features for each target variable, we use LASSO (Tibshirani 1994), a method for estimating generalised linear models using convex penalties (l_1) to achieve sparse solutions (Friedman et al. 2010).

Assume $\bar{t}$ is the mean of the target variable:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

The coefficient of determination R^2 is now defined as:

$$R^2(t, f) = 1 - \frac{\sum_i (t_i - f_i)^2}{\sum_i (t_i - \bar{t})^2} \quad (1)$$

where $\sum_i (t_i - f_i)^2$ is the sum of squared differences between the actual and predicted target value, and $\sum_i (t_i - \bar{t})^2$ is the sum of squared differences between the target value and the constant function $t = \bar{t}$. Note that R^2 is between 0 and 1 whenever the model f is produced using the Ordinary Least Squares method, but may be lower than 0 for functions derived differently. R^2 is often interpreted as the *explained variance*, where a value of 0 means that no variance in the dependent variable can be explained by variance in the independent variable(s), and a value of 1 means that all variance can thus be explained (a perfect fit of the data).

4.2 Subgroup discovery

The previous section focussed on linear models, assuming that the dependencies we hope to discover are indeed linear in nature. Unfortunately, in the domain we are focussing on, it is quite likely that the relationship between (extent of) training and performance is non-linear. When doing a certain training routine, it can be expected that the relationship is in fact curved, with peak performance being achieved at a certain optimal load on the human body. Doing too little will not achieve the right effect, but doing too much of the training also produces sub-optimal performance. Specifically, one can expect thresholds in the training load above (or below) which performance will rapidly diminish. Therefore, we will also experiment with modelling techniques that are more local in nature, and find subsets of the data where performance was surprisingly low, as well as finding variables and thresholds that will identify such sub-optimal subsets.

Our paradigm of choice for such (potentially) non-linear data is that of *Subgroup Discovery* (Klösgen 2002; Novak et al. 2009). It is a data mining framework that aims to find interesting *subgroups* satisfying certain user-specified constraints. In this process, we explore a large search space to find subsets of the data that have a relatively high value for a user-defined quality measure. We consider constraints on attributes, and determine which records satisfy these constraints. These records then form a subgroup. The constraints on the attributes (the *description*) form an intensional specification of a part of our dataset, and the subgroup forms its extension (that is, an exhaustive enumeration of the members of the subgroup).

Subgroup Discovery (SD) is a supervised technique: there is a *target concept* defined on one or several designated target attributes. The most basic form of this is a single target, either nominal or numeric. SD then aims to find subgroups for which the distribution over this target is substantially different from that on the whole population, the extent of which is captured by a *quality measure*. This can be any function $\varphi : 2^D \rightarrow \mathbb{R}$ assigning a numeric value to each subgroup. The choice of quality measure defines what we are looking for in a subgroup.

The specific SD algorithm we use in the experiments uses *beam search* to search through the potentially large space of candidate subgroups. This type of search works top-down, starting with a consideration of all single-condition subgroups. Each candidate subgroup is evaluated on the data by means of the quality measure, and sufficiently interesting subgroups are added to the result set. Of the considered subgroups, only the most promising ones (in terms of quality) are added to the *beam*, which are the subgroups that will be extended with new conditions to form the next level of the search. The size of the beam w is a parameter that is set by the analyst. Smaller values will produce a more greedy search, while larger values will cover a larger portion of the search space. How many levels the search will continue (and hence how complex the subgroup descriptions will be) is governed by the *search depth* parameter d. Especially in our case, where fairly little data is available, d will be set to a fairly small value (at most $d = 3$ in our experiments).

A number of papers in the literature discuss SD variants for regression tasks, which to some extent are applicable to our case. One group of techniques focusses on finding subgroups where the target shows a surprisingly high (or conversely, low) average value (Grosskreutz and Rüping 2009; Atzmüller and Lemmerich 2009; Pieters et al. 2010; Lemmerich et al. 2015). Typical proposed quality measures use statistical tests to capture the level of deviation within the subgroup, often weighted by the size of the subgroup, for example the *mean test* or *z-score* (Pieters et al. 2010),

$$\varphi_z(S) = \sqrt{|S|}\frac{\mu_S - \mu_0}{\sigma_S}$$

where μ_S and μ_0 stand for the subgroup and database means of the target, respectively, and σ_S denotes the standard deviation within the subgroup S. Other works consider the distribution of target values within the subgroup (Jorge et al. 2006), and use statistical measures for assessing distributional differences.

In the majority of these quality measures, the interestingness is computed from the distribution of the subgroup alone, or when compared to that of the entire dataset. Here, we take a slightly different approach, and consider the subgroup description a dichotomy of the data, where both the distribution of the subgroup as well as of the *complement* play a role in determining the quality of the dichotomy. Therefore, we introduce a new quality measure for numeric targets in SD. This quality measure uses the well-known notion of R^2 to capture how well a subgroup and its corresponding complement describe the data, in comparison to the distribution of the entire dataset, so ignoring the dichotomy. Hence, we treat the subgroup as a model of the data, to be more specific a step function of two parts. The following two averages over the target t provide the constant prediction for, respectively, the subgroup and its complement:

$$\bar{t}_{subg} = \frac{1}{|S|}\sum_{i \in S} t_i$$

$$\bar{t}_{comp} = \frac{1}{n - |S|}\sum_{i \notin S} t_i$$

These two average values now lead to the following step function:

$$f_S(i) = \begin{cases} \bar{t}_{subg} & \text{if } i \in S \\ \bar{t}_{comp} & \text{if } i \notin S \end{cases}$$

The quality measure *Explained Variance* is now simply defined as follows:

$$\varphi_{EV}(S) = R^2(t, f_S)$$

This quality measure uses the definition of R^2 given in the previous section, which was independent of the nature of the model f. Note that by using this quality measure, we have a way of directly comparing the discovered subgroups (with corresponding step functions) to the linear models, which is a clear benefit over the traditional quality measures. We furthermore observe that the step functions, despite representing dichotomies, can be based on subgroups of multiple conditions. Therefore, the resulting step functions will be multi-dimensional (involving potentially multiple features). The following proposition states that the values of Explained Variance are bounded by [0, 1]. The usual interpretation of 1 as a perfect fit, and 0 as none of the variance is explained, still holds.

Proposition 1 *Given a subgroup S, then* $\varphi_{EV}(S) \in [0, 1]$.

Proof (*Bounds on regression quality*) For R^2 to be bigger than 1, the quotient (see Eq. 1) would need to result in a negative number, which is impossible since both terms are sums of squares. So, we can focus on proving the $\varphi_{EV}(S) \geq 0$ part. From the definition of R^2, we see that we need to prove that $\sum_i (t_i - \bar{t})^2 \geq \sum_i (t_i - f_i)^2$. Note that

$$\sum_i (t_i - f_i)^2 = \sum_{i \in S} (t_i - \bar{t}_{subg})^2 + \sum_{i \notin S} (t_i - \bar{t}_{comp})^2$$

Differentiation of any sum of squared differences $\sum_i (x_i - c)^2$ will show that it is minimised when $c = \bar{x}$, so

$$\sum_{i \in S} (t_i - \bar{t}_{subg})^2 \leq \sum_{i \in S} (t_i - \bar{t})^2$$

and likewise for the complement. Combining these results, we get

$$\sum_i (t_i - \bar{t})^2 = \sum_{i \in S} (t_i - \bar{t})^2 + \sum_{i \notin S} (t_i - \bar{t})^2 \geq \sum_i (t_i - f_i)^2$$

which proves our premise, and the proposition. □

The quality measure introduced here was implemented in the Cortana Subgroup Discovery tool (Meeng and Knobbe 2011).

Table 1 Excerpt of the training data

Date	*ampm*	*dur*	*int*	*load*	Comment
Tue May 13	am	150	3.5	525	$2 \times (20'D1; 5'D3)+$ mostly D1, little D2
Tue May 13	pm	50	4	200	Slight cold, temp 37.1
Wed May 14	am	60	2.5	150	Temp 36.8, push hip in on left corner
Wed May 14	pm	90	5	450	Blocks $4'50''$
Thu May 15	am	60	3	180	Inline skating, cold
Thu May 15	pm	60	4	240	
–					

5 Experimental results

In order to demonstrate the kinds of analyses and results of the proposed methods on actual data, and to test the benefits of individual techniques proposed above, we experiment with data from four athletes of the LottoNL-Jumbo team, two male and two female. All experiments were performed using three software components:

1. A relational database that organises all the different datasets and meta-data concerning exercise and competition: the *Performance Sports Repository* (PSR).
2. A dashboard accessible over the Internet, that provides various views and visualisations of the data, and allows online aggregation and linear modelling of the data.
3. The generic Subgroup Discovery tool Cortana,[6] which was extended for this purpose with a direct database connection to the PSR, and the Explained Variance quality measure (Meeng and Knobbe 2011).

Table 1 shows an example of the core data considered in these experiments: the training data (in this case for a female skater specialising on the longer distances). Note the diversity of the comments in the last column. These are currently not included in the experiments described below. The column *int*, and consequently *load*, are subjective, meaning they express the perceived intensity of the training. Although this is of course intentional, subjectivity comes with its limitations. It should be noted though that all experiments are athlete-specific, such that differences in how intensities are rated between subjects is not an issue. This training data is then combined with a scraped dataset about competition results, and aggregated to the level of competitions, as described in previous sections.

We will demonstrate the results on the datasets listed in Table 2, collected from four athletes. Next to competition data, we also have physical test data, for which we provide results for one of the athletes (M1 in the table below), for which we have 146 records.

We will generally describe three types of modelling of the data: (1) univariate models, either using a linear or a step function, where we rank all features by R^2, (2) multivariate models using LASSO, and (3) Subgroup Discovery using Cortana. Note

[6] Sources in Java and an executable of this tool can be downloaded from http://datamining.liacs.nl/cortana.html.

Table 2 Dataset details for competition results of four skaters

	Gender	Distance (m)	Competitions	Sessions
M1	Male	1000	75	2758
M2	Male	500	142	2930
F1	Female	1500	60	2230
F2	Female	500	22	1095

that univariate step models can also be interpreted as subgroups with a single condition, such that results between settings 1 and 3 overlap to some extent. When mining for subgroups, we use beam search to a fairly shallow depth, typically to a maximum depth of three or less, depending on the experiment. When not further specified, the subgroups (or their step functions) presented involve a single feature ($d = 1$). The width of the beam is set to a default of $w = 100$ (candidates that proceed to the next level). For the numeric attributes, the Cortana setting 'best' is applied, which means that for each attribute, all numeric threshold are considered and the optimal split point is selected. The resulting locally optimal subgroup is added to the result list if of sufficiently quality, and conditionally added to the beam for further refinement.

Before considering more systematic experiments, for example, testing the merits of the Fitness-Fatigue model, we first present results for a single skater, and demonstrate the kinds of input given to the coach concerning possible changes to the training schedule.

5.1 Demonstration of results

This section discusses results for female skater F1 who specialises in the medium to long distances. We first focus on 1500 m races, for which we have 60 examples over a period of 5 years. The average relative time for this skater was 1.0391, so 3.91% above the track record. We have training details for the entire 5 years, such that we can aggregate the preparation for each of these races easily.

Uniform features We start with univariate results in the simplest setting: uniform aggregates without specifiers and simple linear models. The best-fitting aggregate that was found was `max_load_1` with the following model:

$$y = -0.000014x + 1.042$$

The explained variance of this model is a mere $R^2 = 0.0563$. The model starts with a reasonable intercept, and encourages a high load (the product of duration and intensity) on the day prior to the race. Although the effect is minor, this suggests that a peak right before a race (possibly due to another race in the same weekend) is beneficial. The step function associated with this aggregate function, with a threshold around 360, has a more pronounced $R^2 = 0.1233$. The two levels are $t_{rel} = 1.043$ for low maximum loads versus 1.031.

Table 3 Overview of selected uniform features

Feature	Linear model	R^2 linear	R^2 step	Low	High
`max_load_1`	$-0.000014x + 1.042$	0.0563	0.1233	1.043	1.031
`avg_intensity_19`	$-0.003952x + 1.052$	0.0557	0.2231	1.044	1.028
`max_duration_int5_17`	$0.000211x + 1.029$	0.1491	0.3080	1.035	1.074

The second-best aggregate is `avg_intensity_19`, with model

$$y = -0.003952x + 1.052$$

which suggests that the intensity should be kept high over a period of almost 3 weeks to improve race times. This aggregate actually scores highest on explained variance of the step function, with a minimum average intensity of 3.93 (low to moderate intensity). Lower intensities suggest an expected relative time of $t_{rel} = 1.044$, whereas higher values on average lead to relative times of $t_{rel} = 1.028$ ($R^2 = 0.2231$). For clarity, details of these three features are listed in Table 3.

Specifiers The addition of specifiers does a great deal to the quality of the univariate models. The following aggregate scores the best on R^2:

- `max_duration_int5_17` (the largest period spent at intensity 5 for the last 17 days prior to the race)

Exaggerating this type of training has a detrimental effect on the race outcome: longer durations of this specific training type lead to higher times. The step function ($R^2 = 0.3080$) caps this value at 70 min. With longer durations of intensity 5, relative times of $t_{rel} = 1.074$ are expected, compared to $t_{rel} = 1.035$ below this threshold (Fig. 3).

Fitness-Fatigue model Switching from uniform to FF features, we note that the following parameters provide the best (linear) aggregate, based on the sum of duration: $\tau_{fit} = 39$, $\tau_{del} = 4.0$, $\tau_{fat} = 7.0$, $K = 2.0$. The corresponding kernel is the one featured in Fig. 2 earlier in this paper. The associated explained variance is $R^2 = 0.1002$.

Multi-variate model The individual features presented so far do not lead to very well-fitting models, despite their role in informing the coach of ways to optimise the training and avoiding some pitfalls of under- and overtraining. More precise models can of course be obtained by involving multiple features. The graph in Fig. 4 presents the 60 results achieved by the skater, as well as the predicted times, by a multi-variate linear model. The model, induced by the LASSO procedure, involves 18 features, selected from the larger pool of uniform aggregates (ignoring the specifiers). The quality of the model is $R^2 = 0.721$, obtained on the training set.[7] Although such models (and more accurate models involving more complex aggregates) can be used to predict the outcome of an upcoming race, this particular prediction is of limited value. Rather,

[7] No distinction between training and test set was made for these experiments.

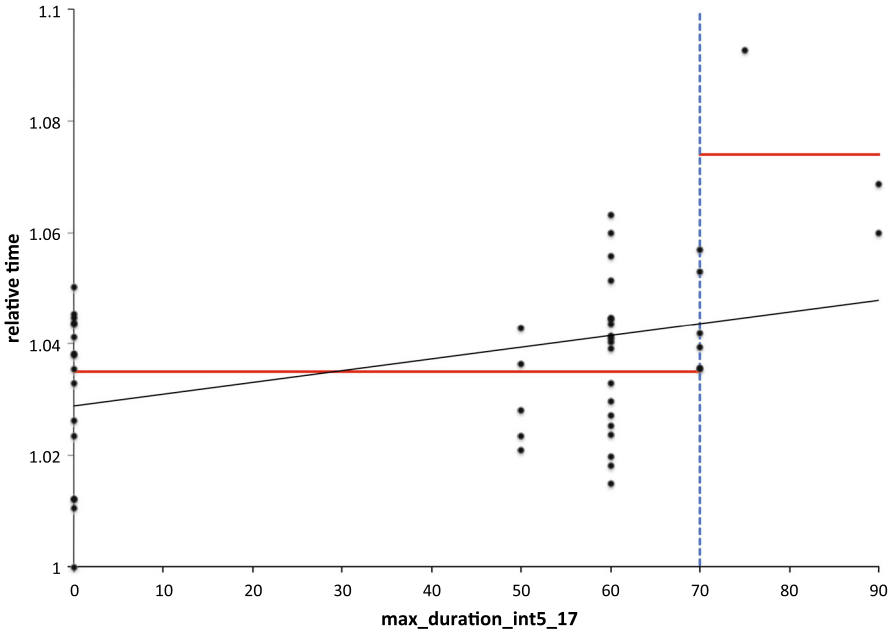

Fig. 3 Graph showing the relation between `max_duration_int5_17` and the relative time of the subsequent race. The *vertical blue line* indicates the threshold (inclusive to the *left*) and the *red* the two average times for subgroup and complement. The *black line* indicates the best-fitting simple linear equation (Color figure online)

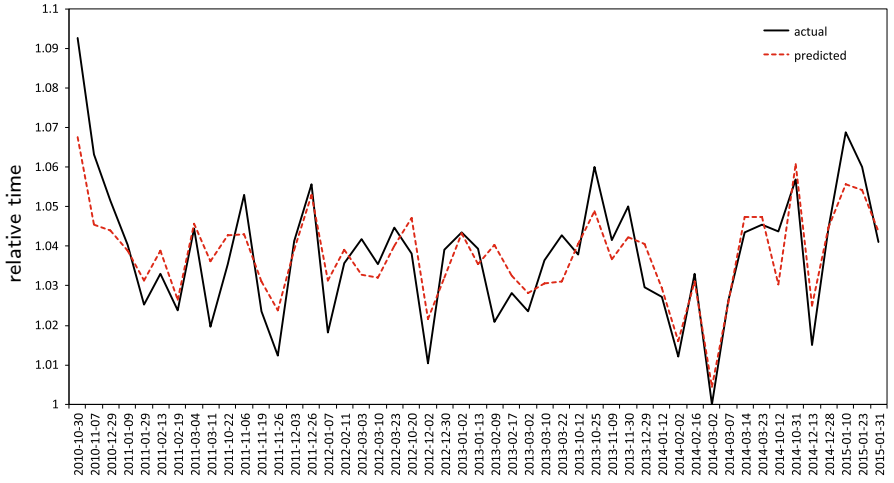

Fig. 4 Obtained results over 60 historical races, compared to the predictions of a relatively simple LASSO model based on the training sessions prior to each race

the model is more valuable from a knowledge discovery point of view, pointing to the features that matter most. In this case, the 18 features mostly include duration over short windows (1–5 days), and intensities over longer windows (approx. 2 weeks).

Subgroups In the above text, we have already presented the following three subgroups[8]:

- $\texttt{max\_load\_1} \leq 360$ ($R^2 = 0.1233$)
- $\texttt{avg\_intensity\_19} > 3.93$ ($R^2 = 0.2231$)
- $\texttt{max\_duration\_int5\_17} \leq 70$ ($R^2 = 0.3080$)

The first two subgroups relate to the experiment with just uniform aggregates, where the second subgroup is the optimal step function found at depth 1. The third subgroup relates to the experiment involving specifiers. While subgroups at depth 1 are interesting since they point to individual predictive features, they capture only shallow effects. We now present subgroups at greater depth, that indeed describe more complex concepts. The best subgroup found on the uniform windows (without specifiers) by Cortana at depth $d \leq 2$ is

$$\texttt{avg\_intensity\_20} > 3.94 \vee \texttt{sum\_duration\_2} > 170 \quad (R^2 = 0.4232)$$

Although Cortana produces subgroups as conjunctions of conditions, for reasons of presentation this was logically inverted[9] in the above subgroup. The subgroup, covering 17 cases, describes races with an average of 1.0299, compared to 1.0526 for the remainder. It specifies that whenever the average intensity over the last 20 days is too low, and the total duration of exercises over the last 2 days is also low, this has a negative effect on the race result. Note how the explained variance has almost doubled at $d \leq 2$. Adding a third feature to the subgroups only produced a marginal improvement, which is not uncommon in SD.

The addition of specifiers in combination with deeper subgroups produced slightly better results, with the top subgroup being as follows:

$$\begin{aligned} &\texttt{avg\_duration\_int5\_17} < 60 \vee \\ &\quad \texttt{sum\_duration\_int6789\_10} > 115 \quad (R^2 = 0.446) \end{aligned}$$

Note that compared to the $d = 1$ result of $R^2 = 0.3080$, this is a reasonable improvement. The subgroup specifies that a lower duration of intensity 5 exercises over 17 day, or a higher duration of high-intensity exercises over 10 days, is good.

Validation of results For the results above, one could wonder to what extent each result is statistically significant. For individual models, be it linear or step function, it is possible to compute a p value that indicates to what extent the model might be due to chance. Such p values will be reported in the detailed experiments in the next section. However, we should note that we are generating a substantial number of features, such that we are in fact testing multiple hypotheses. The best ranked result may thus appear to be significant, even though this is just a consequence of the many models considered.

[8] Although subgroups are interpreted here as dichotomies, we present them as either lower or upper thresholds, such that cases meeting the condition(s) specified correspond to the faster races.

[9] The original subgroup discovered, $\texttt{avg\_intensity\_20} \leq 3.94 \wedge \texttt{sum\_duration\_2} \leq 170$, covers the complement.

Duivesteijn and Knobbe (2011) presents a method for validating the results of an SD algorithm, by means of a *distribution of false discoveries*. This distribution is obtained by running the algorithm repeatedly on the data after swap-randomising the target attribute, thus capturing what maximum qualities can be obtained from random data (that resembles the original data). Using the distribution, it is possible to set a lower bound on the quality (in this case explained variance) as a function of the desired significance level α. Assuming a significance level $\alpha = 0.05$, this validation method produces a lower bound of $R^2_{min} = 0.2907$ for the uniform data without specifiers, searching for subgroups at depth $d = 1$. This means that our optimal subgroup

$$\texttt{avg\_intensity\_19} > 3.93 \;\; (R^2 = 0.2231)$$

is not actually significant at $\alpha = 0.05$. It is good to note that the lower bound produced by the swap randomisation depends on the specific settings of the SD run. Specifically, if the extent of the search is bigger, more hypotheses will be tested, such that the lower bound will increase in order to account for the higher probability of finding a seemingly interesting subgroup by chance. When increasing the search depth to $d \leq 2$, the procedure produces a lower bound of $R^2_{min} = 0.4212$. This makes the earlier depth 2 result (without specifiers)

$$\texttt{avg\_intensity\_20} > 3.94 \vee \texttt{sum\_duration\_2} > 170 \;\; (R^2 = 0.4232)$$

just barely significant.

Åstrand test So far, we have been focussing on the performance at competitions. The LottoNL-Jumbo data also contains physical test results, of which we present one example here. One of the standard ways of assessing an athlete's general fitness is called the *Åstrand test* (Åstrand and Ryhming 1954). It measures the aerobic capacity in terms of a standard value called VO_2max (Kenney et al. 2015; McArdle et al. 2014), and is of relevance to disciplines that require prolonged expenditure of energy. The data we analyse here contains 146 examples of measurements of this value for a specific male skater, nearly twice as many as the number of competition results available for this skater. The average VO_2max for this skater is 59.95 ml/kg/min (higher values indicate better fitness) with a standard deviation of $\sigma = 4.05$. Subgroup Discovery turns out to produce better results than simple linear models, so we focus on that approach.

At depth $d = 1$, the best performing feature (ignoring specifiers), is

$$\texttt{max\_duration\_21}$$

(the length of the longest session over a 3-week period). When this value is above 130 min, the VO_2max is elevated to 61.085 ($R^2 = 0.202$). The fact that duration plays a key role here makes sense: the aerobic system is primarily trained with exercises that last long (as opposed to say sprints or weight training). The fact that the maximum value is apparently relevant is interesting, since it suggests that relatively long sessions

are required, but a single such session may be sufficient. With the same set of features, the following $d = 3$ subgroup raises this value to 62.514 ($R^2 = 0.322$):

$$\begin{aligned} \texttt{max\_duration\_17} &\geq 140 \wedge \\ \texttt{count\_15} &\leq 88 \wedge \\ \texttt{max\_load\_21} &\leq 1250 \end{aligned}$$

Furthermore, with the addition of specifiers (*ampm* and intensity intervals), the following subgroup scored best at $d \leq 3$:

$$\begin{aligned} \texttt{max\_duration\_int123\_17} &\geq 110 \wedge \\ \texttt{sum\_duration\_int456\_7} &\leq 1975 \wedge \\ \texttt{sum\_duration\_int7\_15} &\leq 311.0 \end{aligned}$$

This subgroup, which covers 65 tests (vs. 81 tests in the complement), represents an average of 62.84 ($R^2 = 0.408$). It combines a minimum duration for the longest low-intensity exercise with two upper bounds on the total duration of high-intensity exercises. All Åstrand results presented here are significant at the $\alpha = 0.05$ level, with considerable margin.

5.2 Fitness-Fatigue model

In this section, we analyse the specific merits of the Fitness-Fatigue model introduced in Sect. 3.2. We start by considering the four parameters the model involves, in order to fit the kernel to the specific physiological properties of the individual athlete. Rough values for the optimal setting were determined by informal experimentation, after which an extensive grid search was used to determine the optimal values for each athlete involved. These results are demonstrated in Table 4, for four athletes and their respective distance of speciality. The left columns indicate the gender of the athlete, the preferred distance, the average relative time, and the number of races n available. The remaining columns indicate the optimal values for τ_{fit}, τ_{del}, τ_{fat} and K. The three decay parameters are in days. As an optimisation criterion, we select the R^2 of the (univariate) linear model of the best feature found.

In order to analyse the stability of these parameters, we selected the third athlete (the one with the most available data), and varied each parameter individually, fixing the remaining the parameters to the optimum found earlier. The R^2 of the best feature was recorded for each setting of the parameters. Figure 5 demonstrates for each parameter how sensitive it is to change, in terms of quality of fit of the FF model. We note that all functions are very well-behaved and smooth over the domains considered, with the selected optimum clearly being undisputed. Furthermore, observe that the functions appear to be convex, making them fairly straightforward to optimise. Hence, the relatively simple grid search used in the pragmatic setting can be easily replaced by a more efficient hill-climber.

Table 4 Optimal parameters for the Fitness-Fatigue model for four speed skaters

	Input context				Optimal parameters			
	Gender	Distance (m)	Time	n	τ_{fit}	τ_{del}	τ_{fat}	K
M1	Male	1000	1.0207	75	13	2.0	2.0	2.00
M2	Male	500	1.0212	142	21	2.0	2.0	3.75
F1	Female	1500	1.0391	60	39	4.0	7.0	2.00
F2	Female	500	1.0691	22	29	4.0	4.0	2.00

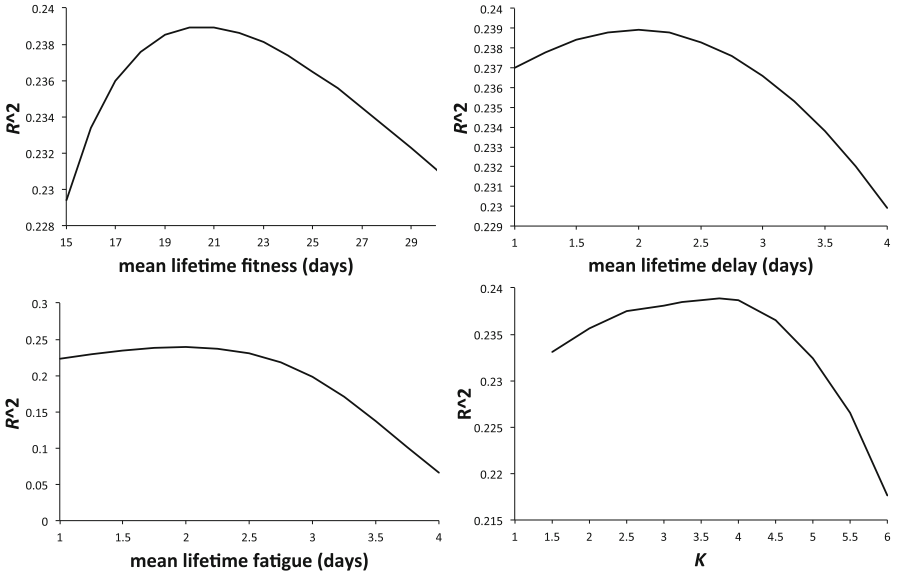

Fig. 5 Analysis of sensitivity of the features to varying the four parameters τ_{fit}, τ_{del}, τ_{fat} and K (from *top left* to *bottom right*). Clearly, these functions are smooth and well-behaved

Let's consider the table of FF parameters in more detail. First of all, the rough numbers are very plausible from a physiological point of view. Clearly, the fatigue and (delayed) gain in fitness should be in the order of a few days, while the prolonged benefit of the exercise remains for a longer period in the order of several weeks. Also, the optimal values clearly differ per athlete, as a function of the different physiology and type of training the athlete is subjected to generally. Table 4 also suggests a difference between men and women, with men having a shorter time scale than women, both for the recuperation and how long the benefit lasts, although such conclusions are hard to draw from only four cases. Note also that the values reported here are somewhat different from the ones reported in Calvert et al. (1976), which are: $\tau_{fit} = 50$, $\tau_{del} = 5$, $\tau_{fat} = 15$, $K = 2.0$. As a last observation, we note that τ_{del} and τ_{fat} tend to assume very similar values. What impact this has from a physiological perspective (the fatigue and beneficial adaptation of the body go hand in hand?) is hard to say, but

Table 5 Comparison of goodness of fit of best univariate linear model for (middle) the uniform aggregates and (right) the Fitness-Fatigue model

	Gender	Distance (m)	Uniform		Fitness-Fatigue	
			R^2 linear	p value	R^2 linear	p value
M1	Male	1000	**0.41**	4.16×10^{-10}	0.30	3.85×10^{-7}
M2	Male	500	0.17	2.46×10^{-7}	**0.23**	7.02×10^{-10}
F1	Female	1500	0.06	9.95×10^{-3}	**0.11**	1.69×10^{-3}
F2	Female	500	**0.54**	9.53×10^{-5}	0.50	2.58×10^{-5}

Bold values indicate the fit of the best model for each skater

at least from a modelling perspective, it is a good opportunity to dispense with one parameter, and make the fitting of models more efficient.

Having a stable and physiologically plausible FF model of the training response, it is now time to turn to the question whether the model indeed produces a better fit, compared to our baseline of uniform aggregates. To this end, we again consider the explained variance R^2 of the linear model on the best feature found, first for uniform features, then for the exponentially decaying features. Since above, the FF model was optimised without specifiers (intensity zones and morning/afternoon distinction), we compare the results with a similar setting for the uniform features. Table 5 presents these results. The columns marked "R^2 linear" indicate the explained variance of the simple linear model. The indicated p values for each result refer to the statistical significance of a linear regression t test: the significance of the best model, testing the hypothesis that the coefficient of the model is not 0 (in other words, testing whether the dependent variable is indeed influenced by the independent variable).

Based on these numbers, there is clearly not a consistent benefit of the FF model, over the less natural uniform features. Especially in the case of the first athlete, the uniform features are in fact more accurate. The feature in question (although there are multiple variants of similar score) concerns the sum of the duration over 9 days, which is an indication of too intense training and hence too high levels of fatigue as a result. Also for the fourth athlete, the uniform features come out on top. Still, for individual athletes the FF model may show a considerable improvement in fit over the unnatural uniform features.

Alternative aggregate functions The presentation of the FF model in terms of convolution translates into SQL as the `SUM` aggregate function. If features based on `SUM` have a potential benefit, so might the alternative functions `AVG` and `MAX`. We present an additional experiment in Table 6 that investigates the added value of these two aggregates to the plain implementation of the FF model used so far. The last two columns of the table show that in two cases, `AVG` or `MAX` do outperform the standard convolution. This is interesting, since there is clearly the potential to improve the models in this way. However, the features are less intuitive to understand since they are not based on the standard definition of convolution in terms of summation.

Table 6 Analysis of the benefit of adding AVG and MAX as aggregates to the Fitness-Fatigue model

	Gender	Distance (m)	SUM		SUM, AVG, MAX	
			R^2 linear	p value	R^2 linear	p value
M1	Male	1000	0.30	3.85×10^{-7}	0.30	3.85×10^{-7}
M2	Male	500	0.24	7.02×10^{-10}	**0.28**	1.60×10^{-11}
F1	Female	1500	0.11	1.69×10^{-3}	**0.15**	5.22×10^{-4}
F2	Female	500	0.50	2.58×10^{-5}	0.50	2.58×10^{-5}

Bold values indicate the fit of the best model for each skater

Table 7 The top six of a ranking of features, ordered by the R^2 of the linear functions

Feature	Linear function		Step function	
	R^2	p value	R^2	p value
sum_duration_int4567_9	0.56	1.207×10^{-14}	0.52	3.974×10^{-13}
sum_duration_int456_9	0.54	5.789×10^{-14}	0.54	8.987×10^{-14}
sum_duration_int456789_9	0.53	1.014×10^{-13}	0.45	3.878×10^{-11}
sum_duration_int4567_11	0.53	1.022×10^{-13}	0.47	9.474×10^{-12}
sum_duration_int45678_9	0.53	1.134×10^{-13}	0.45	3.878×10^{-11}
sum_duration_int34567_9	0.52	2.434×10^{-13}	0.49	2.575×10^{-12}
–	–	–	–	–

The second feature happens to be the optimal one for step functions ($R^2 = 0.5367$). p values of the linear functions are computed as before. The p value for step functions is obtained by performing a Student's t test Rice (2006) on the two samples (subgroup and complement), testing for significant difference between the two means

Table 8 Comparison of best fit between linear models and step functions

	Gender	Distance (m)	Linear function		Step function	
			R^2	p value	R^2	p value
M1	Male	1000	**0.56**	1.21×10^{-14}	0.54	8.99×10^{-14}
M2	Male	500	0.19	7.22×10^{-8}	**0.27**	3.43×10^{-11}
F1	Female	1500	0.19	1.83×10^{-4}	**0.31**	2.63×10^{-5}
F2	Female	500	0.58	4.28×10^{-5}	**0.59**	3.69×10^{-5}

Bold values indicate the fit of the best model for each skater

5.3 Comparing linear and step functions

So far, we have been focussing on aspects of feature construction. In this section, we turn to the challenge of finding regression models within those features, that relate each feature to the target. As discussed in Sect. 4, we consider both simple linear models and step functions as univariate regression models. Thus, we would like to

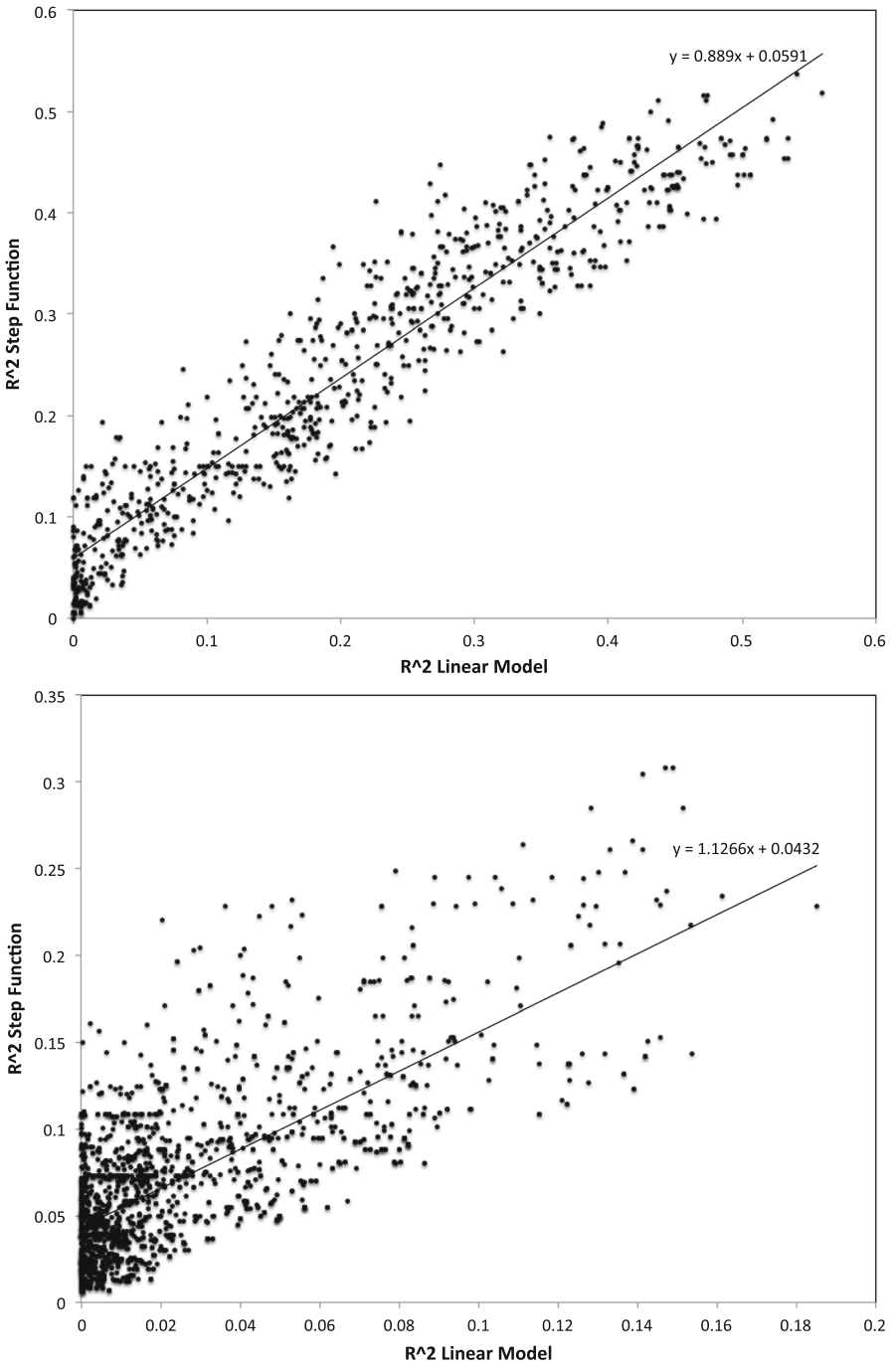

Fig. 6 Plot showing the roughly linear relationship between the R^2 of simple linear functions and the step function derived by Subgroup Discovery. *Each dot* represents a constructed feature and how predictive it is for the target under two conditions: 1000 m for a male speed skater (*top*), and 1500 m for a female skater

know which of these two approaches produces the best fit, and to what extent the approaches rank similar features similarly. Consider the top six features (uniform aggregation, including specifiers) in Table 7 in terms of R^2 of the simple linear model, for male skater M1. Column 2 and 4 show the explained variances, which are clearly very correlated. In fact, the top feature for the step function happens to be the second feature. This suggests that linear models and step functions are roughly as expressive for this athlete.

A more thorough analysis is given in the two plots of Fig. 6. Here, every feature considered is represented by a dot with the R^2 of either model as the x- and y-coordinate. A linear correlation, especially when close to the line $y = x$ suggests that the two models are equally strong. The first plot shows M1, the athlete with the best correlation between the two fits (of the four athletes featured earlier), and the second F1, the athlete where the correlation is the least. The rank correlation for the first two rankings is $\rho = 0.954$, compared to $\rho = 0.646$ for the second. The interesting rules, where linear models fail to fit the data while a step function is able to model the data to some extent, appear on the left of these plots. Interestingly, there are some features that score near zero for the explained linear variance, while still showing some form of dependence using the step function. In Table 8, the R^2 of the best features are listed for each of the athletes. As is clear, in three of the four cases, step functions outperform the linear models, making them a very viable solution to model such data.

6 Conclusion

We have presented a general approach to the modelling of training data in elite sports, with a specific application to speed skating in the LottoNL-Jumbo team. Our approach computes the combined effect of a training schedule by aggregating details of the individual training sessions, and thus capturing a considerable number of aspects of training and how one prepares for important test moments, such as physical tests and races. Since it is not entirely clear from the literature what aspects of training contribute the most, and what parameters individual athletes need to tweak in order to optimise the training to their specific physique and specialisation, we produce a reasonably large collection of promising features. The most relevant features are then selected by a number of techniques, specifically univariate linear regression, the LASSO regression process and Subgroup Discovery. The linear modelling methods assume that the dependencies of interest are indeed monotonic and linear, that is, adding more load to the exercise will increase the (long-term) fitness of the athlete's body. Clearly, this is not generally the case, and one would expect there to be certain thresholds, above which training is no longer beneficial. For each aspect of training, there is an optimal volume, above or below which training is ineffective. This suggests that non-linear models, or models that are able to represent thresholds (such as subgroups) will outperform linear models.

As mentioned in our introduction, we aim to discover interpretable and actionable patterns in the data, such that the coach can immediately incorporate the most significant findings in the preparation for upcoming events, as well as in future training schedules. We believe that our presented approach, that deliberately presents simple

results, and gives clear guidelines and boundaries on training load, makes this possible. In fact, individual findings on the athletes of the team have led to (subtle) modifications in training regimens, most notably where sprinters were sometimes subjected to too much aerobic exercise. It is good to stress again the athlete-specific nature of our analyses. Luckily, for a reasonable number of skaters, we have a long enough history to have a substantial database of training-response examples, where natural variation in preparation has produced a productive dataset. Athlete-specific data leads to athlete-specific findings, and one should therefore not interpret any discovered pattern as a general rule of exercise physiology, but rather as an opportunity to optimise training for that athlete.

We have presented a number of anecdotal results for a specific skater, demonstrating that interpretable and actionable results can be found. The best-fitting subgroup suggests that for a good result, this skater should avoid longer exercises at intensity 5 (over a longer window), as well as (slightly) increase the exposure to intensities 6 to 9. Although separate results appear very significant, a more thorough analysis using swap-randomisation is necessary to account for the many features and models being considered. For this specific skater, statistically significant results could be obtained, despite there only being 60 available races. With up to 142 race results, other skaters will allow for much more significant findings.

The Fitness-Fatigue model, introduced as a more natural way to aggregate training impulses over time, produced reasonable results. After experimenting with four different skaters, two male and two female, very consistent and realistic values were found for the four parameters of this model. Although slight variations did occur, most notably between the male and female skaters, the general picture did match that of the coach. Knowing these (athlete-specific) parameters in detail allows the coach to mix exercise and recuperation in a more precise manner. From a modelling point of view, we also demonstrated that the optimal values of these parameters can be found efficiently, due to their well-behaved nature.

In a number of detailed experiments, we compared different choices in our modelling approach. A first experiment compared the uniform window to the Fitness-Fatigue kernel. Given the unnatural nature of a rectangular kernel, one would expect the FF model to be superior. Somewhat surprisingly, our data did not support this hypothesis. The FF model did indeed produce superior models for two athletes, but the uniform window performed equally superior on the remaining two, leaving this comparison unresolved.

A further experiment considered whether using the aggregate functions `MAX` and `AVG`, alongside the more obvious `SUM`, would be beneficial. This was indeed sometimes the case, although not by a large margin. Whether such slightly more accurate models are in fact attractive is questionable, since the combination with the non-trivial FF kernel does not lead to very interpretable patterns.

Finally, a head-to-head comparison between linear models and step functions was performed. It turned out that in three of the four athletes considered, there is a clear preference for step functions, with explained variance being almost twice as high in one case. This suggests that in these datasets, there are indeed thresholds, above or below which the performance is significantly compromised. Still, we observed that the explained variances between linear and step function are fairly correlated, when

considering all available features. Especially for the one athlete where step functions showed no benefit, the explained variances were very similar, suggesting that the data in question was actually fairly linear. Although in the training domain, one would always expect a U-shaped relation between load and relative time (neither too much nor too little is optimal), this is not necessarily reflected in the data of an athlete's practice. For example, if one consistently trains above the optimal load or intensity, not all parts of the U-curve are being sampled, and available data might appear to be fairly linear or at least monotonic. For at least three skaters in the team, the variance in the training data is sufficient to exhibit the non-linear behaviour that is captured by Subgroup Discovery. We intend to consider more complicated models that specifically aim to capture the non-linear, U-shaped relationships one expects in training data.

The current distribution of relative times (see Fig. 1) generally shows a bell-shaped curve, although the curve has some properties of a long tail on the right. This long tail represents higher race results, where perhaps the athlete made a mistake or gave up altogether. In order to accommodate for this lack of symmetry in the distribution, we could consider alternative distributions such as the *log-normal* or *Weibull* distributions. Such distributions are studied in detail in Survival Analysis (Hosmer et al. 2008), and in future work, we will see if more accurate modelling of this asymmetry can be obtained from inspiration from this field.

Further plans for future research in this field include the idea of analysing data from multiple athletes simultaneously, whether this includes athletes of comparable distances or simply taking data of all athletes available. Developments in the field of multi-task learning (Caruana 1997) could be relevant here. One of the biggest limitations in the analyses so far is the lack of sufficient races for a given athlete and distance. For several athletes in the team, this prevents us from getting significant results. With a combined approach, a general model could be produced on considerably more data, that then needs to be tailored to the individual athlete's situation.

Acknowledgements This research was financially supported by the Netherlands Organisation for Scientific Research NWO, and by the COMMIT\ ICT research community under project *Monitoring and Analysing Speed Skaters*.

Appendix A: Aggregation features in SQL

One of the ways in which the feature construction discussed in Sect. 3.1 can be performed in relational databases is by means of a database view. The below statement assumes there are two tables, one listing all results of races (`Result`), and one listing all exercises (`Exercise`). There is an implicit one-to-many relationship between these two tables, if one assumes that all exercises prior to a race provide information as to the preparation for a race. In order to put a limit on the extent of history that is included for each race (in this case, 4 weeks), there is a

```
DATEDIFF(Result.date, Exercise.date) < 28
```

condition in the where clause, combined with a

```
Exercise.date < Result.date
```

condition to guarantee only involving exercises prior to the race. Since there is a one-to-many relationship, and we are interested in an analysis on the Result level, we need to aggregate the many exercises by means of aggregate functions. The statement below demonstrates five such aggregates, four of which include attributes of the exercise. Note the intensional definition of load as the product of duration and intensity.

```
CREATE VIEW RaceView AS
SELECT Result.id, Result.athlete_id,
    COUNT(*) AS count_28,
    SUM(Exercise.duration) AS sum_duration_28,
    SUM(Exercise.intensity) AS sum_intensity_28,
    SUM(Exercise.duration*Exercise.intensity) AS sum_load_28,
    AVG(Exercise.duration) AS avg_duration_28,
    ...
FROM Result, Exercise
WHERE Exercise.athlete_id = Result.athlete_id AND
    Exercise.date < Result.date AND
    DATEDIFF(Result.date, Exercise.date) < 28 AND
GROUP BY Result.id, Result.athlete_id;
```

The outline below shows features related to the largest window selected, in this case 28 days. Since no windows longer than 28 days are presented to the aggregate functions, no specification of this limit is required. However, the majority of features will involve shorter windows, so for such definitions, the below examples demonstrate how to limit the number of exercises involved to the appropriate window. These four aggregates are the 2-week counterparts of the ones in the above statement. Note the two different values in the then-else clause that depend on the type of aggregate function. For `COUNT`, this needs to be `1` (inside the 2-week window) and `NULL` (outside the window). For `SUM` and `MAX`, this needs to be the attribute in question, and `0`. Finally, for `AVG`, this needs to be the attribute and `NULL`.

```
COUNT(CASE WHEN DATEDIFF(Result.date, Exercise.date) <= 14
    THEN 1 ELSE NULL END) AS count_14,
SUM(CASE WHEN DATEDIFF(Result.date, Exercise.date) <= 14
    THEN Exercise.duration ELSE 0 END) AS sum_duration_14,
AVG(CASE WHEN DATEDIFF(Result.date, Exercise.date) <= 14
    THEN Exercise.duration ELSE NULL END) AS avg_duration_14,
MAX(CASE WHEN DATEDIFF(Result.date, Exercise.date) <= 14
    THEN Exercise.duration ELSE 0 END) AS sum_duration_14,
```

Whenever a specifier is involved in the aggregate, this can be specified with a case-when-then-else clause as well, as demonstrated below:

```
SUM(CASE WHEN Exercise.session = "am" THEN Exercise.duration ELSE 0 END)
    AS sum_duration_am_28,
SUM(CASE WHEN Exercise.intensity >= 4 AND Exercise.intensity <= 6
    THEN Exercise.duration ELSE 0 END) AS sum_int456_28,
```

Appendix B: Fitness-fatigue in SQL

In Sect. 3.1.2, we observed that uniform aggregation using `sum` could be interpreted as convolution with a rectangular kernel, making the implementation in SQL fairly

straightforward. With some minor changes, this also holds for our more sophisticated Fitness-Fatigue kernel. The following aggregate computes the desired feature for the FF kernel, with four parameters supplied.

```
SUM(((EXP(-DATEDIFF(Result.date, Exercise.date)*(1/50)) -
      EXP(-DATEDIFF(Result.date, Exercise.date)*(1/5))) -
     2*EXP(-DATEDIFF(Result.date, Exercise.date)*(1/15))) * Exercise.duration)
    AS FF_duration_50,
```

Note how this aggregate uses the elapsed time (`DATEDIFF`) between the exercise and the competition as the value m in the kernel h_{ff}. Since the kernel asymptotically approaches 0 with increasing time, there should not be a limit on the time between exercise and competition, as is the case with the uniform window (28 days). However, in practical implementations, one can opt for approximate features by limiting the history of each competition, thus gaining efficiency.

References

Atzmüller M, Lemmerich F (2009) Fast subgroup discovery for continuous target concepts. In: ISMIS, the international symposium on methodologies for intelligent systems, pp 35–44

Åstrand P-O, Ryhming I (1954) A nomogram for calculation of aerobic capacity (physical fitness) from pulse rate during sub-maximal work. J Appl Physiol 7(2):218–221

Cachucho R, Meeng M, Vespier U, Nijssen S, Knobbe A (2014) Mining multivariate time series with mixed sampling rates. In: UbiComp, the international joint conference on pervasive and ubiquitous computing, pp 413–423

Calvert TW, Banister EW, Savage MV, Bach T (1976) A systems model of the effects of training on physical performance. IEEE Trans Syst Man Cybern SMC 6(2):94–102

Caruana R (1997) Multi-task learning. Mach Learn 28:41–75

Duivesteijn W, Knobbe A (2011) Exploiting false discoveries—statistical validation of patterns and quality measures in subgroup discovery. In: ICDM, the international conference on data mining, pp 151–160

Hosmer DW, Lemeshow S, May S (2008) Applied survival analysis: regression modeling of time-to-event data. Wiley, New York

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

Grosskreutz H, Rüping S (2009) On subgroup discovery in numerical domains. Data Min Knowl Discov 19(2):210–226

Jorge AM, Azevedo PJ, Pereira F (2006) Distribution rules with numeric attributes of interest. In: PKDD, the European conference on principles and practice of knowledge discovery in databases, pp 247–258

Kenney WL, Wilmore J, Costill D (2015) Physiology of sport and exercise. Human Kinetics, Champaign

Klösgen W (2002) Subgroup discovery, handbook of data mining and knowledge discovery. Oxford University Press, Oxford

Lemmerich F, Atzmueller M, Puppe F (2015) Fast exhaustive subgroup discovery with numerical target concepts. Data Min Knowl Discov 30(3):711–762

McArdle WD, Katch FI, Katch VL (2014) Exercise physiology: nutrition, energy, and human performance. Williams & Wilkins, Lippincott

Meeng M, Knobbe A (2011) Flexible enrichment with Cortana-software demo. In: BeneLearn, the annual Belgian-Dutch conference on machine learning, pp 117–119

Novak PK, Lavrač N, Webb GI (2009) Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J Mach Learn Res 10:377–403

Pieters BFI, Knobbe A, Džeroski S (2010) Subgroup discovery in ranked data, with an application to gene set enrichment. In: Preference learning 2010 at ECML PKDD, the European conference on machine learning and principles and practice of knowledge discovery in databases

Rice JA (2006) Mathematical statistics and data analysis. Duxbury Advanced Series

Stranneby D, Walker W (2004) Digital signal processing and applications. Elsevier, Amsterdam

Tibshirani R (1994) Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B 58:267–288
Vandewalle D, Gilbert P, Monod H (1987) Standard anaerobic exercise tests. Sports Med 4(4):268–289