Universiteit Leiden
The Netherlands

Informatica & Economie

Quantifying and analyzing the load on a day of an athlete

Cas Haasdijk

Supervisors:
Arno Knobbe & Marco Spruit

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/11/2023

Abstract

This research investigates the process of collecting, processing, and analyzing a dataset pertaining to speed skating. The data consisted of 12 athletes of the TalentNED speed skating-team and consisted of time series data. Next to data that was already being collected using questionnaires, an additional questionnaire was created and distributed on a daily basis to gather data about the mental- and physical load of an entire day of an athlete, including other daily activities which are not sport-related. After analyzing the dataset using various analytical methods, the results suggests that the majority of the data demonstrates a linear association. Non-linear models failed to identify substantial non-linear relationships that could offer practical insights. This implies that the linear relationships captured by the Pearson correlation coefficient adequately represent the underlying dynamics, minimizing the need for more complex modeling techniques in this particular context. When opting for a two-method analysis, the use of a linear metric such as the Pearson correlation coefficient alongside with a non-linear method such as explained variance or quadratic fitting seems to be the most effective approach. Future studies should focus on increasing the number of participants, extending the research duration and including match results data for more reliable and objective results.

Contents

1 Introduction

This thesis describes an approach to collect, process and analyze sports data. More specifically, this research focuses on data of young athletes that participate in high-level speed skating. Speed skating is a form of ice skating where competitors try to cover a certain distance as fast as possible on an oval ice rink. There are different disciplines for speed skating, these are long-track speed skating, short-track speed skating and marathon speed skating. I will focus on long-track speed skating, which takes place on a 400 m oval track. Long-track speed skating is often referred to as speed-skating, and likewise I will use the term speed skating to refer to long-track speed skating. Athletes are usually specialized in either the relatively short distances or the relatively long distances, mainly determined by their physical attributes.

This research has been conducted at Dutch sports organization TalentNED[1]. The mission of TalentNED is to prepare young talented Dutch athletes for the highest level of performance. They do so by coaching athletes during their training sessions, but also provide accommodation for the athletes, as well as the purchase and maintenance of required equipment. TalentNED operates in three sports fields, namely mountainbiking, speed skating and road cycling. The data used in this research was provided by twelve athletes of the speed skating-team of TalentNED. The data used in this research was collected during the period April to July 2023. This period is relatively short compared to similar studies, but due to the time available for this bachelor thesis this period could not be extended. However, the short period of data collection allow for results to be faster available, providing possible benefits for the coaches of TalentNED. In particular for athletes who have recently joined the TalentNED team, quickly available results allow for rapid optimization of the training schedule. During a typical speed skating season, no matches take place between April and October. Because of this, this research will only discuss training data. Because of the availability of ice during this period, little to no training sessions take place on the ice. Frequent training sessions during this period consist of alternative training forms such as road cycling, weight training or roller skating.

The aim of this research is to quantify and analyse the load that athletes experience during their entire day. Besides practising sport for part of the day, athletes have all sorts of other commitments such as cooking, laundry and school work. TalentNED was curious to see how the total activities on a day influence the physical and mental load their athletes experience and hopefully gain new insights to optimize their training programmes. Next to the practical use of this research, the academic aim of this research is to make a comparative analysis of the methods that are available for analyzing sports data. After investigating the various analysis methods, I will draw conclusions about the methods that have been used and try to make recommendations for the use of certain analysis methods.

Following up the aim of this research, we can formulate the first research question:

> **RQ1:** *What is the nature of the relationships within sports training data, and how can they be characterized in terms of linearity, quadraticity, or other mathematical models?*

After analyzing the nature of relationships within the given data, I would like to make recommendations that can be useful for future analysis of speed skating data, and can possibly be applied to

[1] https://www.talentned.nl/visie/over-talentned

data of other sport disciplines as well. This leads to the second research question:

> **RQ2:** *What is the most effective and accurate analysis method or combination of analysis methods for examining sports training data?*

For this research, Sport Data Valley[2] has been an important element. This is a non-profit national platform in the Netherlands for analysis and research of sports data. The participants in this research use SportDataValley to upload data of their training sessions and fill in multiple questionnaires on the platform throughout the day. The data that was collected through SportDataValley will be discussed in more detail in section 3.1.1.

Next to this introduction this thesis is structured as follows; Section 2 discussed relevant literature; Section 3 will discuss the data that has been used for this research; Section 4 describes the methods that were used for this research and section 5 will cover the outcome of these methods; Section 6 will interpret the results and discuss the limitations of this research and Section 7 concludes.

[2]https://info.sportdatavalley.nl

2 Background

2.1 Data collection in sports data science

In the evaluation of parameters pertaining to the health of athletes, a commonly employed methodology involves the utilization of questionnaires. The construction of such instruments entails careful considerations. To optimize athlete participation, questionnaires are typically formulated as briefly as possible, thereby minimizing the time required for completion. Kerr et al. [KHA18] make several recommendations for the execution of research involving athletes, underscoring the imperative nature of comprehensive planning prior to questionnaire deployment. Questionnaires should be designed in a specific way, rather than being an opportunistic practice of data collection. To uphold objectivity, questions are frequently subjected to numerical rating scales. Illustrative inquiries include: "What is the extent of soreness experienced today?" or "On a scale from 1 to 10, what is your level of satisfaction of the past training session?" Discrepancies in questionnaire frequency for athletes exist, with daily administration prevalent in numerous sports disciplines to accommodate the variable nature of daily training sessions. While this approach ensures daily data availability, a potential drawback arises from the increased likelihood of athletes forgetting to complete a questionnaire due to the daily recurrence.

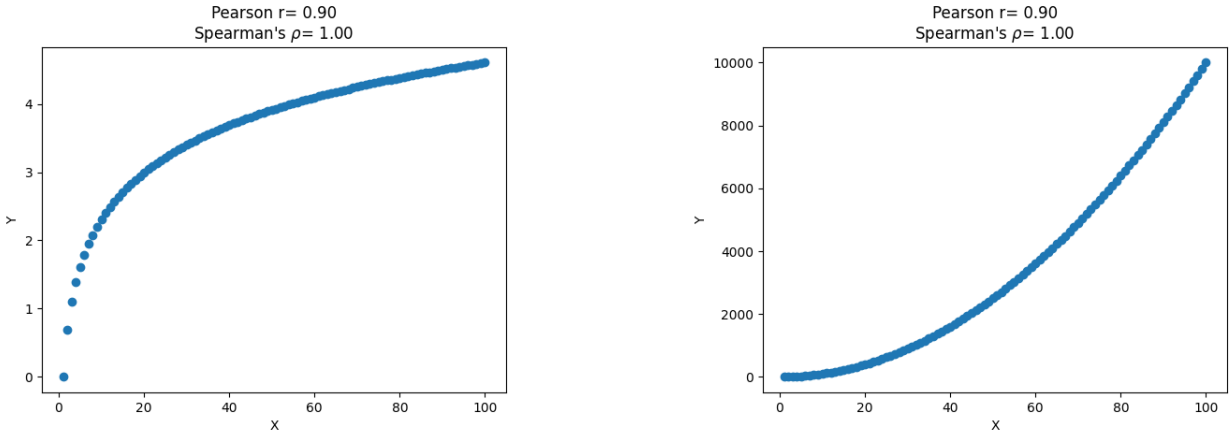
Numerous studies have been conducted to assess not only the physical strain but also the psychological burden experienced by athletes. Mellalieu et al. [MJW+21] discuss the importance of considering the psychological load that athletes experience, as well as the different methods of measurement for psychological load and the challenges that arise when trying to measure the psychological load. They distinguish between 'sport load' and 'life load', where the 'sport load' refers to the total psychological demand an athlete faces from their sport and the 'life load' refers to the total psychological demand an athlete faces away from their sport.

2.2 Univariate data analysis

This research will cover comprehensive analysis of a dataset. When performing data analysis on a set of variables, commonly referred to as features, a target variable is often defined. This is the variable of which the researcher wants to gain a deeper understanding, and in the case of predicting models this variable is the variable that is being predicted by using the other variables. Univariate analysis is a method of analysis in which each variable in the dataset is explored separately. In the case of a target variable, univariate analysis usually covers the target variable and one additional variable. During this univariate analysis, one examines the relationships between one variable and the target variable. Correlation coefficients [Ako18] emerge as a commonly utilized approach for exploring associations within a dataset. One way to measure the linear correlation between two sets of data is the Pearson correlation coefficient. This coefficient represents the strength of a linear relationship between two variables and is denoted as a number between -1 and +1.

Next to linear relationships, datasets may also contain relationships that follow a non-linear trend. Spearman's rank correlation coefficient provides a method for quantifying such non-linear relationships between variables. Instead of looking at a linear relationship, the Spearman's rank correlation coefficient assesses how well the relationship between variables can be described using

a monotonic function. A monotonic function is a function which is either entirely non-increasing or non-decreasing. This coefficient is also denoted as a number between -1 and +1. To elucidate the efficacy of Spearman's rank correlation coefficient, Figure 1 illustrates its values for a natural logarithmic function and a quadratic function in comparison to the values of the Pearson correlation coefficient. Despite a relatively high value for the Pearson correlation coefficient, Spearman's rank correlation coefficient excels in capturing these non-linear relationships, yielding a value of 1.



(a) Coefficients for a logarithmic function

(b) Coefficients for a quadratic function

Figure 1: Values for Pearson's correlation coefficient and Spearman's rank correlation coefficient for a logarithmic and quadratic function.

2.3 Subgroup discovery

I expect some relationships in the dataset to follow a pattern that can not be identified using the previously described methods. For example, match performance is expected to be maximized at a certain amount of training load in preparation of the match event. When this amount of training is not achieved, match performance will be sub-optimal, but overtraining in preparation of a match will also ensure failure to achieve the optimal result. This phenomenon can not be captured using relatively simple models like correlation coefficients. Therefore, I explore the use of Subgroup Discovery. Subgroup Discovery is a data mining technique and attempts to search relations between different properties or variables of a set with respect to a target variable. Subgroup Discovery aims at identifying interesting groups of data points, where 'interestingness' is defined as distributional unusualness with respect to a certain property of interest [Atz15]. To evaluate each subgroup, a quality measure is used [MK21]. A quality measure is a function that quantifies how interesting a subgroup is. In principle, the goal of subgroup discovery algorithms is to identify subgroups that score high on a particular quality measure. A subgroup is a subset of the entire dataset and is usually denoted by its coverage and corresponding conditions for the subgroup. Conditions specifying the characteristics of the subgroup are articulated as a set of constraints that the subgroup must fulfill. Possible conditions are for example a certain variable having substantially lower values than the rest of the data, or a particular variable having substantially lower values than the rest of the data. Subgroup discovery provides a large range of applications. It can be utilized for predictive

analysis such as classification tasks and regression, or it can be used to explore a given dataset. This research uses subgroup discovery, more specifically for exploring a given dataset.

2.4 Analyzing the effects of training load

Boressen and Lambert [BL09] discuss the process of quantifying training load and analyzing the effect of training load on the performance of athletes. In their opinion the most reliable way to measure training load is the heart rate-based method. By monitoring the heart-rate of athletes during their training sessions one can obtain accurate information about the training intensity. However, this entails the necessity to use heart rate monitors throughout training. An alternative way to assess the training load, is to use the session RPE. The session RPE has been introduced by Foster et al. [FDH+96] and is based on the *Rate of Perceived Exertion (RPE)* introduced by Gunnar Borg [BN74]. The RPE scale is a way of measuring physical activity intensity level. This rate is usually scored on a scale ranging from 1 to 10, with 1 indicating little no no effort and 10 indicating maximum effort. The session RPE can be obtained by multiplying the RPE by the session duration in minutes.

Knobbe et. al [KOH+17] researched various aspects of speed skating data. By analyzing training and match data of speed skaters they provide possible ways to tailor the training schedule to the individual athlete. Overall, data analysis in optimizing training schedules for professional speed skaters enables a more scientific and data-driven approach to training, leading to improved performance and reduced risk of injury. In this research, linear modelling and subgroup discovery were used. I will follow a similar approach in this research but will use additional other metrics and will focus on the comparison between these analysis methods.
As discussed, athletes usually specialise in one of the disciplines, divided into sprint, medium and long distance. The physiology of athletes strongly determines the type of distances at which they perform best. For each discipline, a specific type of training is required. Moreover, training for a specific distance can actually negatively impact the performance on other distances [KWC21] [MKK10].

2.5 Time series analysis

Data can occur in multiple forms, for example numerical data, categorical data, or text data. One form that is often handled in sports data science, is time series data. This is data that is collected through repeated measurements over time [Mad07]. Examples of time series data are a currency exchange rate, the number of monthly airline passengers and the daily body temperature of an athlete.

In the examination of time series data, numerous specialized methodologies are frequently employed. Knobbe et al. [KOH+17] delineate these techniques within the context of speed skating time series analysis. Among these methods, the translation of data points emerges as a prevalent approach. This technique involves the systematic displacement of the entire time series along the temporal axis, commonly denoted as "time shifting." The application of time shifting facilitates an exploration of lagged relationships and patterns. In the realm of sports data analysis, this method proves valuable for scrutinizing the impacts of specific parameters. For instance, a sports data scientist

might investigate the effect of a high training load on subsequent soreness levels two days later, accomplished by aligning data from previous days in the same row. Furthermore, the dissection of time series data into discrete subsets is achievable through the use of time windows. A time window is characterized by a specific temporal interval defined by a starting and ending point. Within a given time window, all events falling within its temporal bounds can be aggregated utilizing various aggregation functions. For instance, aggregation methods such as summation or averaging may be applied to the events occurring within the designated time window. These aggregates provide insights into the cumulative effects of multiple events within each time window, thereby condensing disparate data points into a singular representative value.

3 Data

This section will discuss the data that has been used for this research. I will address the collection of the data, then explain how the final dataset was created and discuss the data quality. After establishing the final dataset, exploratory data analysis was performed to glean preliminary insights.

3.1 Data collection

The dataset was constructed by combining data from the SportDataValley platform and by data that was collected using an additional daily questionnaire.

3.1.1 SportDataValley

The data that was available on the Sport Data Valley platform originates from three sources:

- Daily questionnaire filled in every morning by athletes
- Daily wearables data
- Training logs filled in by athletes after each training session

The details of this data will be discussed in the next section.

3.1.2 Daily load questionnaire

To expand the data that was already available on the SportDataValley platform, a questionnaire was designed to try to capture the daily load that athletes experience on a day. The questionnaires that athletes were already filling in are mostly focused on the training sessions and the effects of these training sessions. This newly introduced questionnaire tried to broaden the perspective of the athletes in order to gather information of the load of an entire day. This was emphasized by stating that both training sessions and all other activities on a day should be considered while filling in the questionnaire. The questionnaire has been sent daily to the athletes by Whatsapp and this message was delivered every night at 21:00 for the duration of this research. This time was chosen because at this time, athletes have finished their day for the most part and have enough time to fill in the questionnaire before going to bed.

The questionnaire was designed to gain information about the mental, physical and total load the athletes experience. It consisted of three questions, these were:

1. What is the amount of mental load that you experienced today?
2. What is the amount of physical load that you experienced today?
3. What is the amount of total load that you experienced today?

The questionnaire was created using Microsoft Forms and the results were exported to an Excel file, before being converted to a CSV file and combined with the SportDataValley data. Details of the data that was collected with this questionnaire will be discussed in section 3.2.1. The design of the final questionnaire can be seen in figure 2. Because of the primary language of the athletes, the questionnaire is written in Dutch.

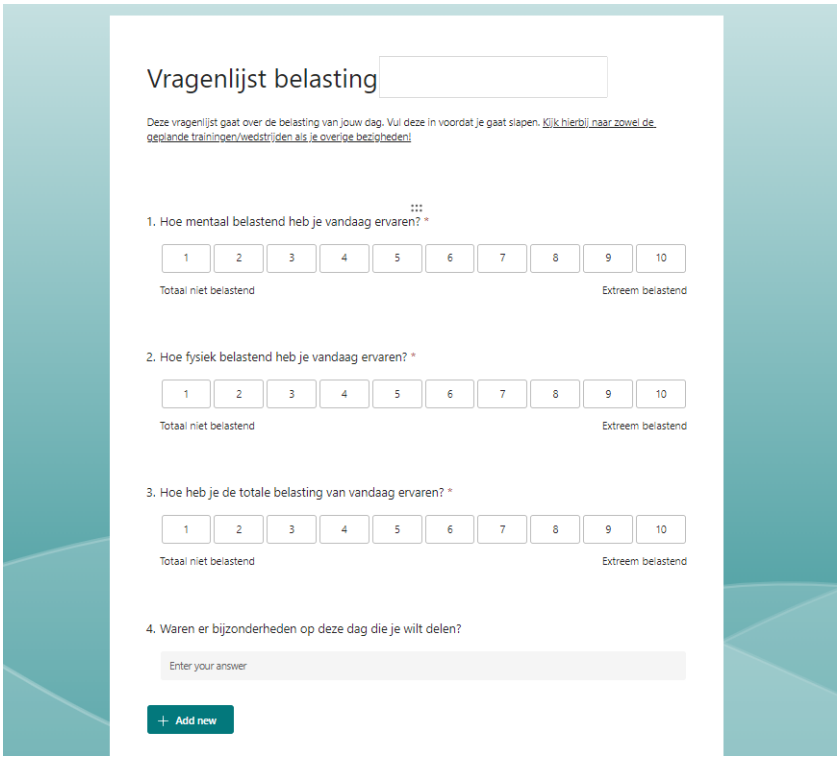


Figure 2: The designed questionnaire

3.2 Data explanation

In this section I will discuss the data that was used in this research. The raw data was cleaned and transformed in order to create additional features.

3.2.1 Raw data

As discussed in sections 3.1.1 and 3.1.2 the raw data originates from two sources. The training logs consist of the following variables:

- **Owner id.** Each individual athlete is identifiable by its owner id, which is a three- to four-digit number.
- **Train start date.** The date of the day on which the training was started.
- **Train duration.** The duration of the training session expressed in amount of minutes.
- **Train RPE.** The rate of perceived effort (RPE) that the athlete filled in after the training session. This is rated on a scale from 1 to 10.
- **Train load.** The train load is expressed as a session-RPE. The session-RPE can be obtained by multiplying the RPE by the duration of the training session.

The data from the daily morning questionnaire consists of the following variables:

- **Daily sleep quality**. The daily sleep quality is rated on a subjective scale ranging from 1 to 5.
- **Daily sleep duration.** The sleep duration is expressed in the amount of hours and filled in by the athlete.
- **Daily fatigue.** The amount of fatigue is rated on a subjective scale ranging from 1 to 5. The higher the score, the less perceived amount of fatigue.
- **Daily stress.** The amount of stress is rated on a subjective scale ranging from 1 to 5. The higher the score, the less perceived amount of stress.
- **Daily soreness.** The amount of soreness is rated on a subjective scale ranging from 1 to 5. The higher the score, the lower the degree of soreness.
- **Daily mood.** The daily mood is rated on a subjective scale ranging from 1 to 5. The higher the score, the better the mood of the athlete.
- **Daily readiness.** The daily readiness indicates the readiness to train and is rated on a subjective scale ranging from 1 to 5. The higher the score, the more ready the athlete feels.
- **Daily weight.** The weight of the athlete expressed in amount of kilograms.
- **Date.** The corresponding date of the filled in questionnaire.

The data that from the wearables consist of the following variables:

- **Resting heart rate.** The daily resting heart rate, expressed in beats per minute.
- **Steps.** The daily number of steps.
- **Calories.** The daily amount of burned calories, as calculated by the wearable.

The data that is available through the daily load questionnaire consist of the following variables:

- **Mental load.** This expresses the mental load that the athlete experienced over the past day.
- **Physical load.** This expresses the physical load that the athlete experienced over the past day.
- **Total load.** This expresses the total load that the athlete experience over the past day.

These three variables are all subjective scores and are rated on a scale from 1 to 10, with 1 indicating an extremely low load and 10 an extremely high load.

3.2.2 Excluded features

Some of the features in the raw data were excluded. These are the number of steps, amount of calories burnt and the daily weight. When importing data of the steps and calories, several inconsistencies were encountered. The data from the wearables contained multiple values for these variables for a single day, making it difficult to obtain reliable data. Next to that, the steps taken on a day do not constitute a good representation about the activity of the athlete throughout the day. While some of the training sessions of the athletes that participated in this research consist of running, most of the training sessions during the research did not involve steps. Because of this, the number of steps does not provide useful information and was excluded during preprocessing of the data. Finally, the data about the daily weight were incomplete and considering the scope of this research the weight does not contain valuable information.

3.2.3 Total dataset

The resulting dataset after selection and exclusion of the variables is shown in table 1. The total data consists of time series data where each of the variables is available for every day in the period of the research.

Variable	Variable description
owner_id	Each individual athlete is identifiable by its owner id
daily_sleep_quality	Daily sleep quality is rated on a subjective scale ranging from 1 to 5
daily_sleep_duration	Daily sleep duration is expressed in amount of hours
daily_fatigue	The amount of fatigue is rated on a subjective scale ranging from 1 to 5
daily_stress	The amount of stress is rated on a subjective scale ranging from 1 to 5
daily_soreness	The amount of soreness is rated on a subjective scale ranging from 1 to 5
daily_mood	The daily mood is rated on a subjective scale ranging from 1 to 5
daily_readiness	Daily readiness to train and is rated on a subjective scale ranging from 1 to 5
resting_hr	The daily resting heart rate, expressed in beats per minute
daily_load	Total of the session RPE's on a day
avg_rpe	The average RPE of the training sessions on a day
max_rpe	The maximum RPE of the training sessions on a day
nr_sessions	The number of training sessions on a day
mental	the mental load that the athlete experienced over the past day
physical	This expresses the physical load that the athlete experienced over the past day
total	This expresses the total load that the athlete experience over the past day

Table 1: Total list of variables available in the dataset and their descriptions

3.2.4 Feature construction

In order to construct additional features, both time shift translation and aggregates were used. The reason for the use of time shifting, is that events of a certain day usually influence the days that follow as well. For example, the training load on day t might influence the readiness to train on day $t + 1$. To investigate these effects, information about previous days was stored in the same

row. In this way, every row contains information about the day itself, but also about previous days. An overview of the time-shifted variables can be seen in table 2. These time-shifted variables were added as a new column to the dataframe and were named to the variable with the suffix `_t-{t}`, where `t` is the period of time-shift that was applied. For example, time-shifting the `daily_load` variable 1 period leads to a new variable called `daily_load_t-1`.

Variable	Periods shifted
daily_load	1, 2, 3
avg_rpe	1, 2, 3
max_rpe	1, 2, 3
mental	1, 2, 3
daily_sleep_quality	1, 2
daily_sleep_duration	1, 2
daily_stress	1, 2

Table 2: Overview of the time-shifted variables

For the construction of aggregates, time windows of the data were used. These time windows ranged from 2 to 5, depending on the variable for which the aggregate is used. An overview of the aggregates that were created can be seen in figure 3. Similar to the time-shifted variables, the aggregates are named to the variables that is aggregated with either the prefix `avg_` or `max_` and the suffix `_t{t}` where `t` is the length of the time window that is used. For example, using the mean aggregate for the variable `mental` with a time window of 3 leads to a new variable called `avg_mental_t3`.

Variable	Aggregate	Time windows
daily_load	mean, max	2, 3
mental	mean, max	3, 5
daily_sleep_quality	mean	3, 5
daily_sleep_duration	mean	3, 5
daily_stress	mean	3, 5

Table 3: Overview of the aggregated variables

3.3 Subjects and data selection

In total, twelve athletes participated in this research. They are all part of the speed-skating team of TalentNED. However, they do differ in the discipline they focus on. This also has slight consequences for their training schedule, which is tailored to the different disciplines of the athletes. The details of the athletes that participated in the research are shown in table 4. Besides their characteristics, the total data points per collection method is also displayed. For each collection method, the maximum possible data points is also noted.
I wish to draw attention to a specific instance concerning the amount of data points pertaining to the resting heart rate for athlete 1629. This athlete used two distinct measurement methods

Owner id	Discipline	Gender	Training logs (max. 138)	Load questionnaires (max. 98)	Morning questionnaires (max. 98)	Heart rates (max. 98)
1818	Allround/long	Male	120	98	96	97
2577	Allround/long	Female	60	97	96	44
511	Allround/long	Male	124	98	96	90
1629	Allround/long	Male	101	97	96	82
1819	Allround/long	Female	91	98	93	59
1319	Allround/mid	Male	108	97	96	75
2573	Allround/mid	Male	123	98	96	78
2583	Allround/mid	Female	120	98	96	60
2572	Allround/mid	Male	74	97	88	78
1817	Allround/mid	Female	104	98	97	87
1821	Allround/short	Male	122	98	96	98
2575	Allround/short	Male	120	98	96	97

Table 4: Information about the subjects that participated in the research

for assessing their resting heart rate, leading to disparate outcomes. Starting from May 2023, athlete 1629 transitioned from manual counting of beats per minute to utilizing a Garmin watch, resulting in a clear reduction in the daily heart rate, as illustrated in figure 3. It is assumed that this alteration does not signify an authentic change in resting heart rate but is rather related to the method of measurement. Consequently, the data from May onwards was considered as reliable, and the remaining data points were omitted from analysis.

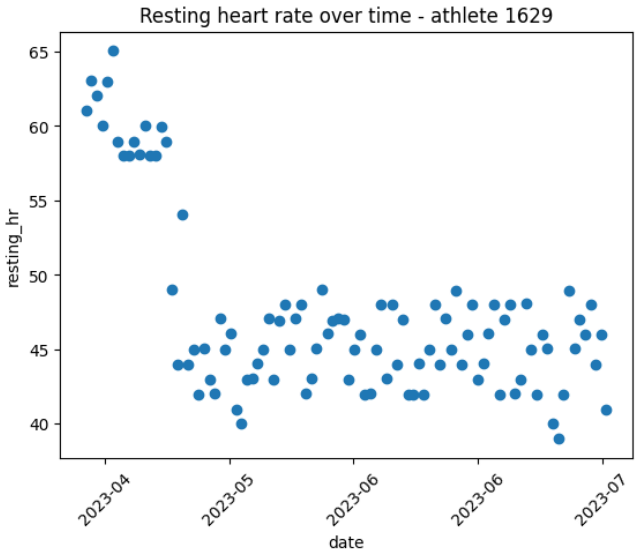


Figure 3: Resting heart rate over time for athlete 1629

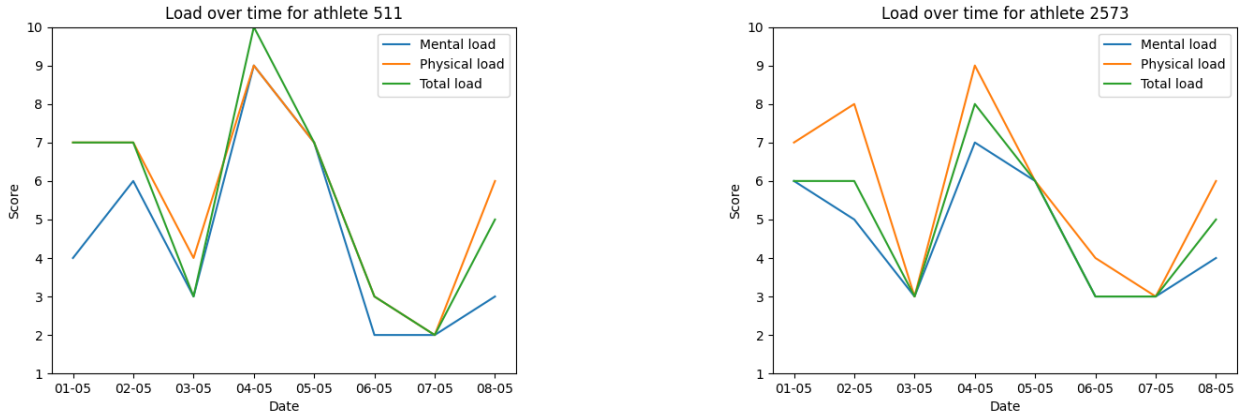
The quality of the data heavily relies on the consistency of filling in the daily questionnaires and training logs of the athletes. Since the amount of data points for the dependent variables is

considered of crucial importance, a minimum amount of 90% is used for the dependent variables. As a result, athletes who filled in less than 88.2 morning questionnaires were excluded in the analysis where daily readiness was used at target variable. Next to that, the influence of training logs has an indirect, but also important effect on the reliability of the data. Data of athletes who filled in less than 70% (96.6) of the total training logs was considered unreliable, so athletes 2577, 1819 and 2572 were removed from the total dataset. In analyses where the resting heart rate was used as target variables, the minimum of 90% was applied so only athletes 1818, 511, 1629, 1821 and 2575 were taken into account.

3.4 Exploratory data analysis

After constructing the total dataset, exploratory data analysis was performed to gain understanding of the dataset and find practical insights for the coaches and athletes of TalentNED. A list of examples will be discussed in this section.

Athletes of TalentNED receive a weekly report including various variables such as the readiness the train, the amount of sleep, and other variables regarding their overall health. The results of the questionnaire that was created for this research were added to this report so athletes and coaches could see how the mental, physical and total load is experienced by the athletes in a week. Examples of the graphs that were created for these weekly reports can be seen in figure 4.



(a) Athlete 511

(b) Athlete 2573

Figure 4: The experienced load in a week for athletes 511 and 2573

Furthermore, I looked at a subset of the data, the rest days. The purpose of a rest day is to allow the athlete to recover from the training sessions of the previous days. To allow for recovery, the physical load on a rest day must be significantly lower than training days. The mean score of both rest days and training days were calculated to compare these scores with each other. The results for the mental load scores are shown in figure 5 and the results for the physical load scores can be seen in figure 6. As expected the experienced physical load is lower on rest days for every athlete. This decrease in load is less visible for the mental load. For one athlete, athlete 1818, the mental load is even higher on rest days than on training days. Although this does not have a direct effect on

recovery, coaches might aim to minimize this mental load on days where athletes have to recover.

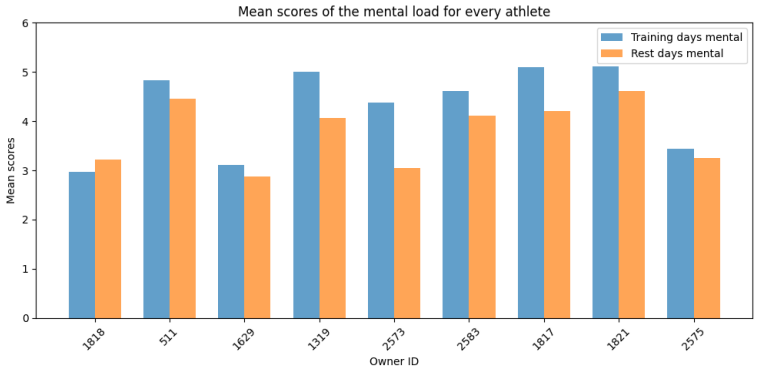


Figure 5: Analysis of rest days versus training days for mental load

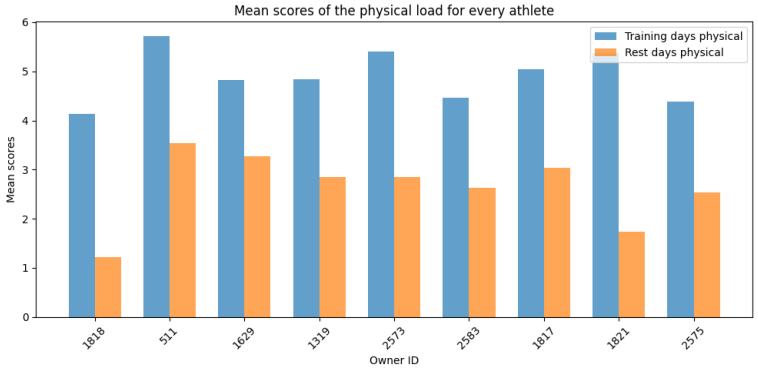
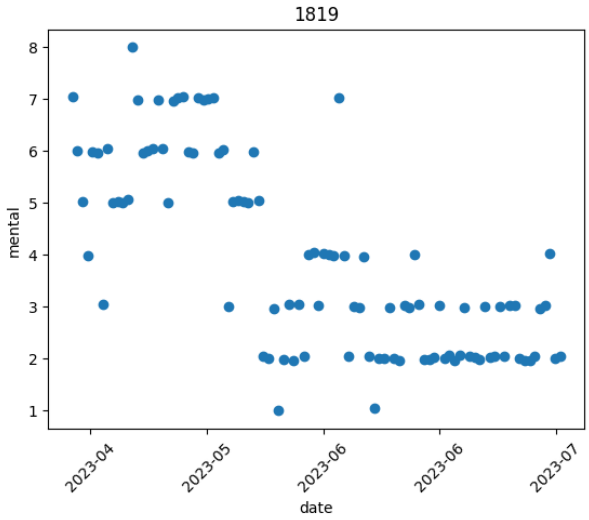
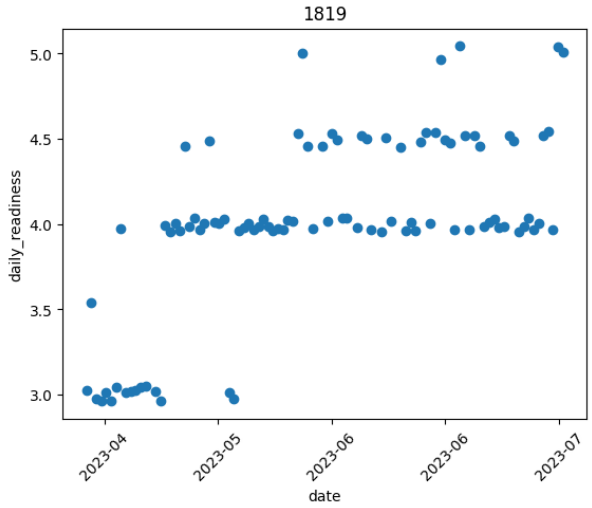


Figure 6: Analysis of rest days versus training days for physical load

For athlete 1819, there was a shift visible over time in both the daily readiness and the mental load experiences. This is shown in figure 7. A possible and plausible explanation for this shift is the period of final exams in the Netherlands, which take place in May every year. Athlete 1819 had exams during this month and probably spent a lot of time and energy in the months before May preparing for these exams. This preparation and the exams themselves may have caused the relatively high mental load and relatively low daily readiness in the months April and May.



(a) Mental load over time for athlete 1819



(b) Daily readiness over time for athlete 1819

Figure 7: The mental load and daily readiness over time for athlete 1819

4 Methodology

This section will discuss the methods that were used during this research. The methods can be separated into three categories. First of all, individual analysis was done to understand the relationship between each of the available variables and different target variables. Consequently, an analysis was performed to investigate whether athletes show similarities in the variables that influence the target variable. Finally, a comparative analysis was performed for the different analysis methods to assess the value of each individual analysis method that was used. A flowchart of the analyses that were performed can be seen in figure 8. The results that were gathered through individual analysis will be utilized to compare individuals with each other, as well as making a comparison between the different analysis methods possible.

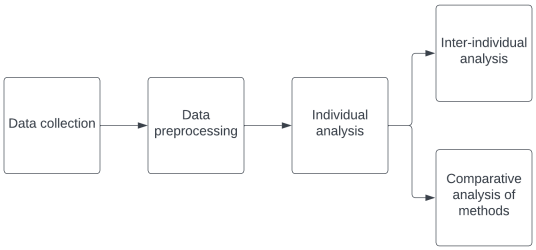


Figure 8: Flowchart of the analyses of the data

Variables originating from the same domain as the target variable were systematically excluded from the dataset. This procedural step was undertaken to mitigate the potential impact of interdependence among independent variables. For instance, in the case of using daily readiness as the target variable, all variables concurrently rated by athletes in the morning questionnaire were excluded from the analytical framework. This criterion is consistently applied across all methods encompassed within the three discussed categories.

4.1 Individual analysis

Several studies have indicated the importance of individual analysis within the domain of sports data science [DHMVY+22]. Instead of examining an entire group of athletes and generalising the data, individual analysis is a way to make personal recommendations based on the characteristics of an individual athlete. This personal approach was also highly emphasized during this research. Several methods have been used to do individual analyses for the athletes. These will be discussed in the next subsections.

4.1.1 Pearson correlation coefficient

The Pearson correlation coefficient is a correlation coefficient that measures the linear correlation between two sets of data. The coefficient ranges from -1 to 1. The Pearson correlation coefficient was implemented by using the `scipy.stats` module[3]. Specifically, the module's `scipy.stats.pearsonr`

[3]https://docs.scipy.org/doc/scipy/reference/stats.html

function was employed, taking `x` and `y` as parameters, both being array-like inputs. The output of the function is a result object which has two attributes; The Pearson correlation coefficient and its associated p-value. This p-value was based on a two-sided alternative hypothesis.

I was interested in calculating the Pearson correlation coefficient between each of the variables and the target variable. Two target variables were used: the daily readiness and the resting heart rate. After defining the target variable, the Pearson correlation coeffcient was iteratively calculated by calling the `pearsonr` function. This function was systematically applied to each combination of a column in the dataset and the target column. The outcome of this process yielded a coefficient for each variable. To gain insight in the influence of all variables, a ranking was created based on the resulting dataframe. The variables were sorted based on the value of the Pearson correlation coefficient. Variables that are positioned higher up in the ranking represent a stronger linear relationship with the target variable.

4.1.2 Spearman's rank correlation coefficient

In addition to examining linear relationships using the Pearson correlation coefficient, this research was extended to the exploration of non-linear associations. Informed by pertinent literature [PM18], I anticipated the existence of non-linear interactions among certain variables. To address this aspect, the Spearman's rank correlation coefficient was employed on the original dataset. This coefficient, operating on the ranks of values, facilitated the analysis of non-linear relationships.

Similar to the Pearson correlation coefficient, the application of the Spearman's rank correlation coefficient yielded distinct coefficients for each variable. An illustrative instance where the Spearman's rank correlation coefficient is expected to outperform the Pearson correlation coefficient relates to the relationship between the training load from the preceding day and the resting heart rate on the current day.

4.1.3 Subgroup discovery

As discussed in the background section, subgroup discovery is a data-mining technique to identify interesting groups of individuals within a dataset. For this research, subgroup discovery was used to try to find non-linear patterns that could not be identified using the Pearson correlation coefficient. The metric chosen to quantify the impact of variables in this process was Explained Variance. The explained variance[4] measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given dataset. This metric makes it possible to identify groups in the dataset which show high variation in comparison to the entire dataset.

Subgroup discovery was implemented using the `pySubDisc` Python wrapper for the Data Mining tool `SubDisc`. Comparable to the method used for the Pearson correlation coefficient, the value for explained variance was calculated for each of the variables. The subgroup discovery wrapper offers several settings. The subgroup discovery wrapper incorporates diverse settings; however, given the objective of comparing variables, the focus was directed towards identifying the single subgroup with the highest quality for each variable. Consequently, the 'best' setting for the `numericStrategy`

[4] https://en.wikipedia.org/wiki/Explained_variation

variable was adopted, ensuring the output exclusively comprised the most optimal subgroup when multiple subgroups were identified. The subgroup tool also provides a way to compute the threshold for the found subgroups. This can be done by using the function `sd.computeThreshold`. By using this function it was possible to calculate the quality level at which the subgroup could be seen as significant and use this quality level as a minimum. The threshold was calculated by performing swap randomization [GMMT07] 100 times. Part of the code that was used to perform the subgroup discovery can be found in Listing 1.

```
sd = pySubDisc.singleNumericTarget(data, target)
sd.qualityMeasure = 'Explained Variance'
sd.qualityMeasureMinimum = -100
sd.numericOperatorSetting = 'LEQ'
sd.numericStrategy = 'best'
sd.computeThreshold(significanceLevel=0.05, method='swap-randomization',
amount=100, setAsMinimum=True, verbose=False)

df_SD = sd.run(verbose=False).asDataFrame()
```

Listing 1: Performing subgroup discovery in Python

4.1.4 Quadratic fitting

Next to the previously mentioned methods, I also used curve fitting to try to find quadratic relationships between the variables and the target variable. This was implemented in Python using the `scipy.optimize.curve_fit` function from the `scipy` module. The model function that was used for quadratic fitting is the following:

$$f(x, a, b, c) = a * x^2 + b * x + c$$

The `curve_fit` function takes the model function, the x-data, and y-data as parameters and uses this to determine the optimal parameters for a, b, c for the model function. It outputs these parameters and its corresponding expected covariance. To assess the goodness of fit for the fitted curve, the R^2 was calculated. The R^2 [5], or Coefficient of determination, is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). This coefficient usually ranges from 0 to 1. This was done by using the `sklearn.metric.r2_score` function.

4.2 Comparing athletes

I was interested to see whether certain athletes show similarities in the way they react to certain training load and other factors that were captured in the features. To compare the athletes that participated in this research, I compared the ranking of variables that was the result of the analysis using the Pearson correlation coefficient.

[5] https://en.wikipedia.org/wiki/Coefficient_of_determination

4.2.1 Spearman's rank correlation coefficient

After analysing the data of the athletes using the Pearson correlation coefficient, this resulted in a ranking of variables based on the values of the Pearson correlation coefficients. To compare these rankings, I used the Spearman's rank correlation coefficient. It assesses how well a monotonic function can be used to describe the relationship between datasets. It can also be described as the Pearson correlation coefficient of the ranks of the variables. The Spearman's rank correlation coefficient was calculated between the rankings for each combination of two athletes and placed in a matrix. It is important to understand the difference between using the Spearman's rank correlation coefficient for individual analysis, which has been described in section 4.1.2, and using Spearman's rank correlation coefficient for comparing athletes. For individual analysis, Spearman's rank correlation coefficient was calculated on the primary data for each of the available variables. This involves all data points of a certain variable and the target variable. For comparing athletes, we take the ranking of variables that was the result of individual analysis. The length of the ranking is equal to the number of variables that are available in the original dataset, minus one because we exclude the target variable in the ranking. We can then calculate the Spearman's rank correlation coefficient for the rankings of two athletes, resulting in one number for each combination of two athletes. These numbers were then converted to a heatmap with a colour scale ranging from blue to red, where blue shows high negative correlation and red shows high positive correlation. In this way I was able to identify athletes that show similarity with other athletes.

4.2.2 Network graphs

Based on the Spearman's rank correlation coefficient, network graphs were made to visualize the degree of similarity between the athletes. Each athlete forms a node in the graph and lines were drawn between each of the nodes. Both colour and thickness of the line indicate the value of the Spearman's rank correlation coefficient between the athletes. Increased line thickness signifies higher coefficients, while the line color follows an ascending scale from blue to red. A red line indicates a high value for the Spearman's rank correlation coefficient, and therefore indicates a high degree of similarity between the rankings of variables of the athletes.

For the implementation of network graphs, the Python module `NetworkX` was used. For the weight of the edges, a simple conditional statement was created which is based on the value of the Spearman's rank correlation coefficient. Since this coefficient can also be negative, the `NetworkX` cannot determine which weight to use for these negative values. So for negative values, a weight of 0.1 is used and for all other values the weight is obtained by multiplying the Spearman's rank correlation coefficient by 10. Part of the code that was used for creating the network graphs is shown in Listing 2.

```
# Choosing the layout algorithm
pos = nx.spring_layout(G, seed=42)

# Create a ScalarMappable object to map colors to correlation values
norm = Normalize(vmin=-1, vmax=1)
cmap = cm.RdYlBu_r
sm = ScalarMappable(cmap=cmap, norm=norm)

# Draw the network graph with varying edge size and color
edges = G.edges(data=True)
edge_widths = [abs(edge[2]['weight']) * 10 if edge[2]['weight'] >= 0
else 0.1 for edge in edges]
edge_colors = [sm.to_rgba(edge[2]['weight']) for edge in edges]

nx.draw(G, pos, with_labels=True, font_size=10, node_size=500,
node_color='lightblue', edgelist=edges, edge_color=edge_colors,
width=edge_widths)
```

Listing 2: Creating the network graphs in Python

4.3 Comparing analysis methods

An analysis was made to compare the different analysis methods that were used during the research. To summarize, the following analysis methods were used:

- Pearson correlation coefficient
- Spearman's rank correlation coefficient
- Quadratic fitting
- Explained Variance (by using Subgroup discovery)

To be able to compare these measures, the Pearson correlation coefficient and the Spearman's rank correlation coefficient were squared to obtain a R score. The R^2 score of the quadratic fitting was calculated using the `sklearn.metric.r2_score` function. The explained variance is a measure that expresses the proportion to which the used model accounts for the variation in a given dataset. Since the R^2 can be interpreted as the explained variance, all discussed measures can be compared with each other. The total result of the comparison of these methods is a table where each of the columns represents an analysis method.

For each athlete there are four rankings, one for each analysis method. To compare these rankings, the Spearman's rank correlation coefficient was calculated. As discussed earlier, this metric considers the ranking of the values instead of the exact values. By calculating the coefficient pairwise between each of the analysis methods, six coefficients were calculated. Each of these coefficients expresses to what degree one analysis method agrees with another method. This resulted in six numbers for each athlete. For each comparison of two analysis methods, the median was taken from the values of all athletes. Using the median ensures a balanced perspective by considering the middle-ranking values rather than being overly influenced by extreme correlations. This methods offers a reliable summary of the degree of agreement between each pair of analysis methods. To visualize these numbers, a matrix was created where the correlation between the analysis methods can easily be seen.

An important note for this comparison between analysis methods is that only the significant variables were considered when calculating the correlation coefficient between the rankings. The variables that were significant for at least one of the methods were taken into account for the comparison. For the Pearson and Spearman's rank correlation coefficients, this was the case for variables with a corresponding p-value smaller than 0.05. For the quadratic fitting, the significance of the parameter a was calculated by using a t-test. This outputs a p-value as well, making it possible to identify variables exhibiting a significant quadratic effect. For subgroup discovery with explained variance as measure, a quality threshold was calculated using swap randomization. Variables with a quality score surpassing this threshold were deemed significant for the purpose of comparison.

5 Results

In this section the results of the applied methods will be shown and discussed. The results will follow the same structure as the methodology section.

5.1 Individual analysis

In this section the results of the individual analyses will be discussed. Although this individual analysis is executed for every athlete, I will not discuss every individual set of results. Instead, a generic approach will be shown to demonstrate how this method can be used for an individual athlete. The results of a number of individuals will be covered.

5.1.1 Pearson correlation coefficient

The use of the Pearson correlation coefficient resulted in a ranking of each of the available variables and was ranked in decreasing order based on the value of the coefficients. An example of the significant results that were obtained for athlete 1818 by using the Pearson correlation coefficient is shown in table 5. The bold p-values are significant at the level $\alpha = 0.05$. As an example to understand these results, the scatterplot for the variable `max_load_t2` is shown in figure 9. Mukkaka [Muk12] provides a rule of thumb for interpreting the value of the Pearson correlation coefficient which is visualized in table 6. Applying this rule of thumb to the results in table 5, we can see that the first variable in the ranking shows a moderate positive correlation with the target variable, while the other variables show a low or negligible correlation.

Variable	Pearson	P-value
max_load_t2	-0.506	**0.00000**
avg_load_t2	-0.498	**0.00000**
max_load_t3	-0.481	**0.00000**
avg_load_t3	-0.476	**0.00000**
daily_load_t-1	-0.399	**0.00015**
avg_rpe_t-1	-0.355	**0.00085**
max_rpe_t-1	-0.353	**0.00093**
daily_load_t-2	-0.317	**0.00308**
avg_rpe_t-2	-0.216	**0.04668**

Table 5: Significant results for the Pearson correlation coefficient for athlete 1818 with daily readiness as target variable

In the available dataset we can distinguish two types of variables. On one hand we can identify aggregate variables, and on the other hand, we applied time shifting to construct variables. I was interested whether there were differences between these two types in terms of usability. Therefore the number of significant variables that were found using the Pearson correlation coefficient have been outlined in Table 7. This figure includes all variables of the analyses of both target variables that were used, both daily readiness to train and the daily resting heart rate. In total, 60 aggregated variables showed a significant linear relationship with the target variables, and 61 time-shifted variables showed a significant linear relationship with the target variables.

Size of correlation	Interpretation
0.90 to 1.00 (-.90 to 1.00)	Very high positive (negative) correlation
0.70 to 0.90 (-0.70 to -0.90)	High positive (negative) correlation
0.50 to 0.70 (-0.50 to -0.70)	Moderate positive (negative) correlation
0.30 to 0.50 (-0.30 to -0.50)	Low positive (negative) correlation
0.00 to 0.30 (0.00 to -0.30)	Negligible correlation

Table 6: Interpretation of the correlation coefficient

Athlete	Aggregates	Time shifts
1818	13	13
511	8	10
1629	12	12
1319	6	4
2573	2	2
2583	0	0
1817	8	4
1821	4	6
2575	7	10
Total	60	61

Table 7: Amount of significant variables for each type of variable

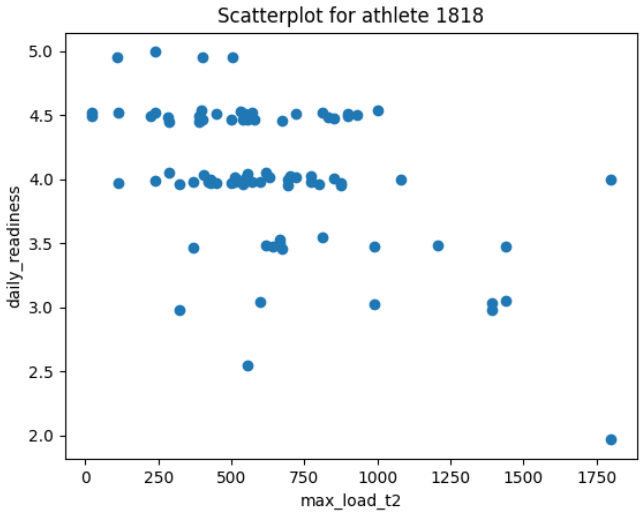


Figure 9: Scatterplot for athlete 1818 which shows the relationship between the `max_load_t2` and the daily readiness to train. `max_load_t2` represents the maximum daily training load of the previous two days.

5.1.2 Subgroup discovery

Performing subgroup discovery resulted in a list of subgroups with a quality value above the computed threshold. This computed threshold serves as a minimum for a subgroup to be significant.

As an example, the results for athlete 1817 can be seen in table 8. The subgroups that are shown in this figure can be considered significant since they have a quality score greater than the calculated quality minimum. Next to the quality, the coverage, average, standard deviation and conditions of the subgroup are given. The coverage of the subgroup refers to the number of data points that the subgroup contains. The conditions tell the properties of each subgroup. So for the subgroup with the highest quality, this subgroup contains 74 data points and each of those data points has a value for `avg_mental_t5` which is lower or equal to 6.0. This implies that when the average of the mental load for the past five days is minimized, this leads to a higher readiness to train. Coaches might use this insight to prepare athletes optimally towards an important event such as a match.

Daily readiness average = 3.814

Depth	Coverage	Quality	Average	St. Dev.	Conditions
1	74	0.259	3.946	0.465	avg_mental_t5 $\leq$ 6.0
1	51	0.172	4.020	0.424	mental_t-2 $\leq$ 5.0
1	79	0.170	3.899	0.496	avg_mental_t3 $\leq$ 6.6666665
1	11	0.148	3.227	0.684	max_load_t3 $\leq$ 200.0

Table 8: Subgroup discovery results for athlete 1817. Significant subgroups are displayed with the corresponding coverage, quality score, average, standard deviation and conditions. Subgroups are ranked on quality score in descending order.

Subgroup discovery was also performed with search depths of 2 and 3. While this increased the number of subgroups that were identified and the explained variance of the highest ranked subgroups, these results are not included in this analysis. Because of the relatively small sample size of the dataset, we estimate the the risk of overfitting quite plausible and therefore limit the analysis to univariate analysis.

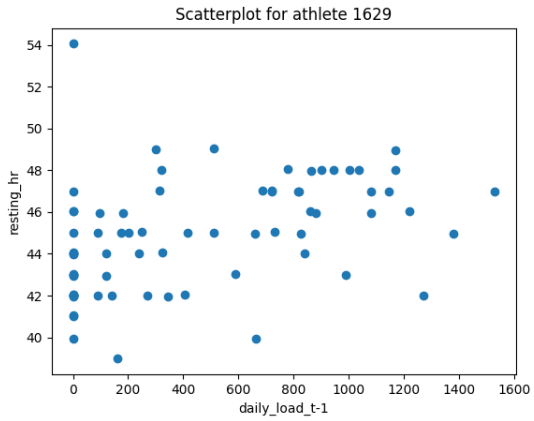
5.1.3 Spearman's rank correlation coefficient

Similar to the Pearson correlation coefficient, I have created a ranking of variables for the values of Spearman's rank correlation coefficient. An example of the list of significant variables for athlete 511 can be seen in table 9. The variable with the strongest correlation is `daily_load_t-1`. This variable expresses the training load of the previous day.

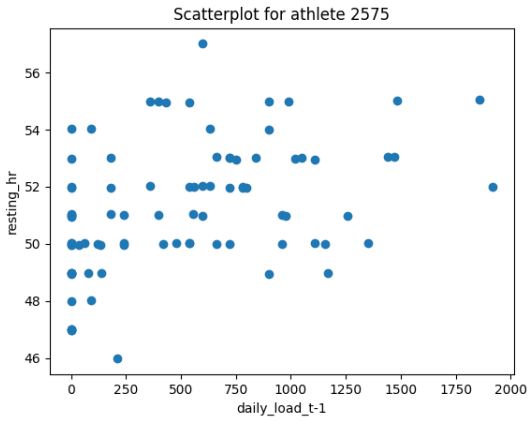
Variable	Spearman	P-value
daily_load_t-1	0.144	**0.00051**
avg_daily_sleep_duration_t5	0.143	**0.00054**
max_rpe_t-1	0.127	**0.00117**
avg_rpe_t-1	0.122	**0.00151**
avg_load_t2	0.070	**0.01751**
nr_sessions_t-1	0.060	**0.02916**
avg_daily_sleep_duration_t3	0.059	**0.03017**

Table 9: Significant variables for Spearman's rank correlation coefficient for athlete 511 with daily readiness as target variable

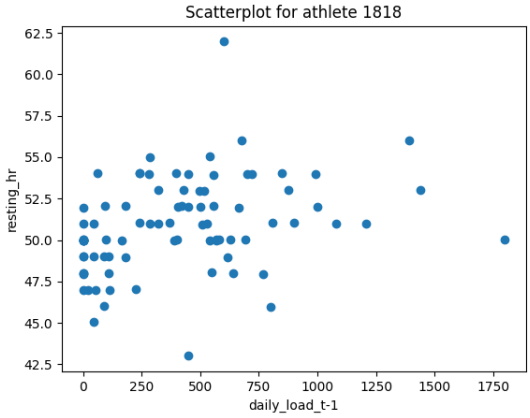
Certain variables exhibited higher Spearman's rank correlation coefficients compared to linear methods such as the Pearson correlation coefficient. One of those variables was the training load of yesterday when considering the resting heart rate as target variable. Figure 10 depicts the scatterplot illustrating the relationship between these two variables for selected athletes. For this variable, the value of the Spearman's rank correlation coefficient versus the value of the Pearson correlation coefficient is shown in table 10. In most cases these two coefficients show very similar scores, although minimal difference can be seen for athlete 1629.



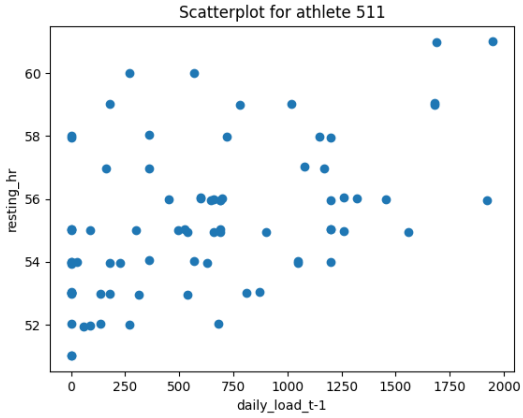
(a) Athlete 1629



(b) Athlete 2575



(c) Athlete 1818



(d) Athlete 511

Figure 10: Scatterplots of the training load versus the resting heart rate for four athletes

Athlete	Pearson	Spearman
1818	**0.332 (p = 0.00194)**	**0.399 (p = 0.00015)**
511	**0.496 (p = 0.00000)**	**0.498 (p = 0.00000)**
1629	**0.423 (p = 0.00017)**	**0.513 (p = 0.00000)**
1821	-0.0655 (p = 0.54897)	-0.0249 (p = 0.82023)
2575	**0.414 (p = 0.00008)**	**0.434 (p = 0.00003)**

Table 10: Comparison between Pearson and Spearman's rank correlation coefficient for relationship between daily_load_t-1 and the resting heart rate. Significant values are displayed in bold.

5.1.4 Quadratic fitting

Quadratic fitting procedures were conducted individually for each athlete, yielding an associated R^2-score for each variable. For individual analyses, these scores were sorted in descending order to create a ranking where the variable with the highest score was on the first position in the ranking. The results for athlete 1319 are shown in Table 11. Although the R^2 is a good measure for the total proportion of variation that is predictable, it does not provide information about the significance of the quadratic parameter. For this reason, a p-value of the a parameter has been calculated with a t-test. This p-value is also shown in Table 11.
An example of a variable which showed a relatively high R^2-score for athlete 1818 and a significant p-value for the a parameter is `daily_load_t-2`. The corresponding plot can be seen in figure 11. The orange line represents the quadratic function that fits the data best. This line was drawn by plotting the quadratic function with the optimal parameters that were provided by the `curve_fit` function. Next to the quadratic model function, a linear function is also added and displayed as `f1` to illustrate the difference between a linear curve fit and a quadratic curve fit. The confidence interval for the quadratic function is shown as well as a grey surface around the quadratic function. This finding suggests that for this athlete, the daily training load of two days earlier should be in a certain range, in order to maximize the readiness to train on the current day. An increasing training load leads to a higher readiness to train until a point where an increase in training load will decrease the readiness to train. For this athlete, this turning point appears to be around a training load of 1250.

Variable	R^2	P-value
max_load_t2	0.264	0.345
avg_load_t2	0.249	0.881
max_load_t3	0.232	0.822
avg_load_t3	0.231	0.500
daily_load_t-2	0.201	**0.002**
daily_load_t-1	0.160	0.805
avg_rpe_t-1	0.126	0.912
max_rpe_t-1	0.125	0.802
avg_daily_sleep_duration_t5	0.064	**0.044**
mental_t-2	0.064	**0.023**

Table 11: Quadratic fitting for athlete 1818 with daily readiness as target variable. For each variable, the R^2 and p-value for the quadratic `a` parameter are displayed

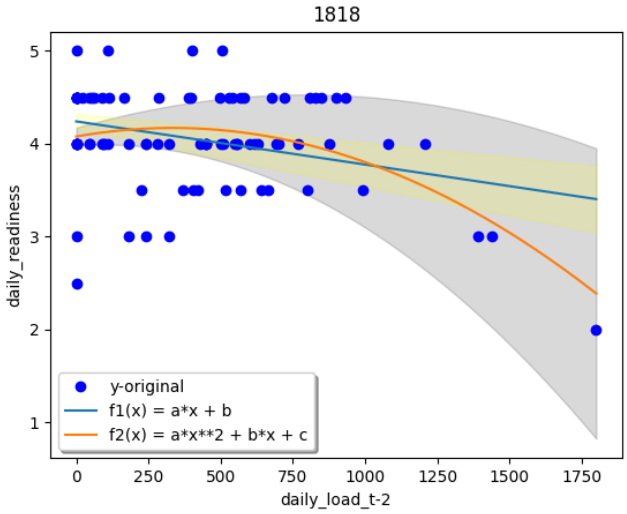


Figure 11: Quadratic fit for athlete 1818

5.2 Comparing athletes

The results of the inter-individual analysis will be discussed in this section.

5.2.1 Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient was employed and visualized through a heatmap to elucidate the degree of similarity between athletes. Exemplifying this approach, Figure 12 illustrates the heatmap derived from the Pearson correlation coefficient, with daily readiness as the target variable. Based on this matrix, it appears that two clusters of athletes can be identified. First of all, athletes 1821 and 2575 show a correlation coefficient of 0.66, indicating a moderate positive correlation in accordance with the classification provided in table 6. Next to that, athletes 1629, 511, and 1818 show correlation coefficients ranging from 0.48 to 0.79. It is relevant to mention that these pairs of athletes share the same gender, and additionally, athletes 1821 and 2575 share a

common discipline.

A parallel analysis was conducted with the resting heart rate as the target variable, and the corresponding outcomes are depicted in Figure 13. In this context, no apparent correlation is evident concerning the discipline of the athletes. Remarkably, among all pairs of athletes exhibiting a correlation exceeding 0.50, only athletes 1818 and 1629 partake in the same discipline.

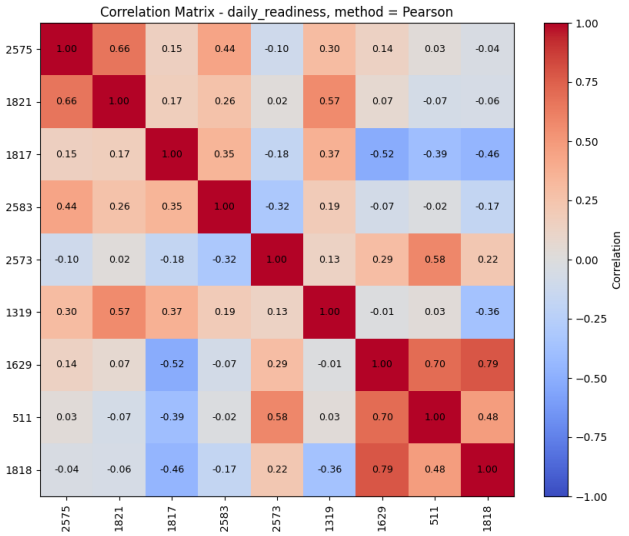


Figure 12: Correlation matrix representing the Spearman's rank correlation coefficient between athletes for the results of the Pearson correlation coefficients with daily readiness as target variable

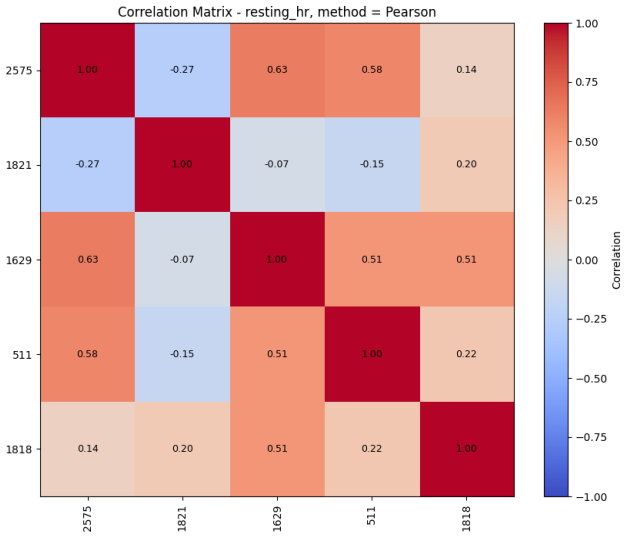


Figure 13: Correlation matrix representing the Spearman's rank correlation coefficient between athletes for the results of the Pearson correlation coefficients with resting heart rate as target variable

5.2.2 Network graphs

The network graphs were made for several target variables and analysis methods. The network graph for the Pearson correlation coefficients with daily readiness as target variable can be seen in figure 14. The network graph where the resting heart rate was used as target variable can be seen in figure 15. The thickness and colour of the edges indicate the degree of similarity between athletes.

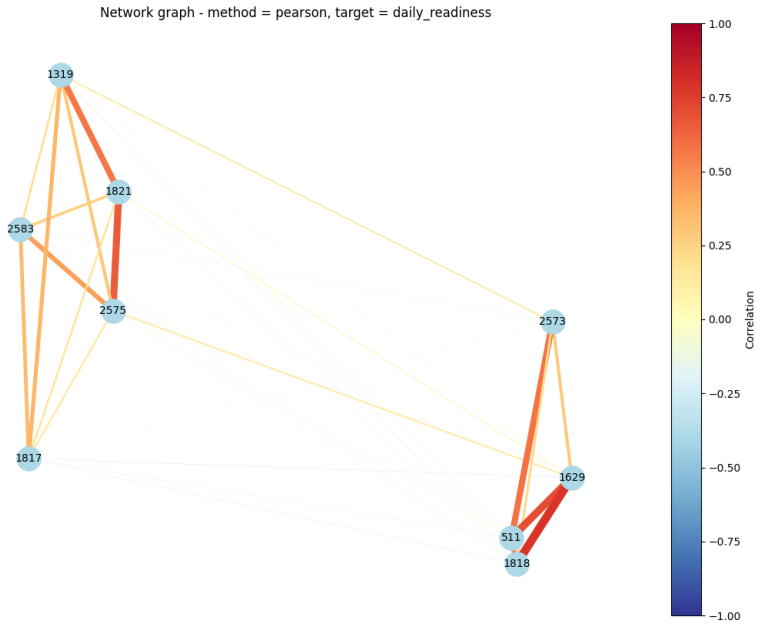


Figure 14: Network graph with daily readiness as target variable

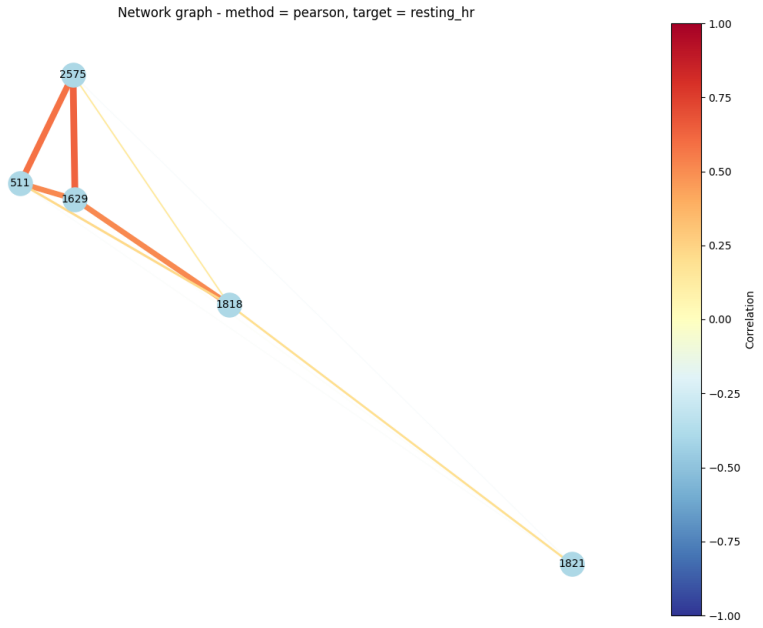


Figure 15: Network graph with resting heart rate as target variable

5.3 Comparing analysis methods

For atlete 1821, a visual comparison between two analytical methods is demonstrated in Figure 16. This representation illustrates the contrast between the utilization of quadratic fitting and the Pearson correlation coefficient. Each data point within the figure corresponds to a variable, with the rank of the variable depicted on the x-axis for one method and on the y-axis for the other method. Red data points signify variables that lack significance for both methods. Notably, this figure reveals a noteworthy convergence in the rankings of variables between the two methods, particularly highlighting the similarity in the scoring of the highest-ranked variables for this specific case. This indicates that both methods yield comparable results and diminishes the rationale for employing both methods. In this instance, the utilization of either method suffices.

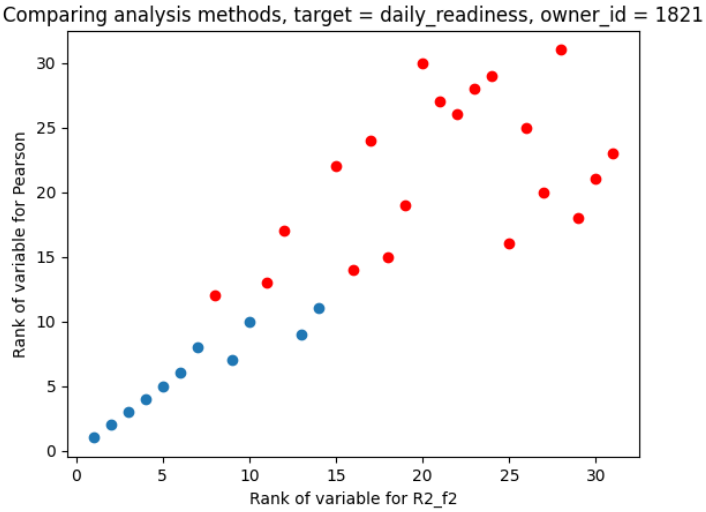


Figure 16: Ranking of variables of the quadratic model versus the Pearson correlation coefficient

For a comprehensive overview of the comparative performance of all methods, a matrix was constructed following the outlined methodology. Figure 17 presents the outcome of this matrix, focusing on daily readiness to train as the target variable. Similarly, Figure 18 showcases the resultant matrix, with the resting heart rate as the target variable. The intensity of coloration in each box within the matrix denotes the strength of the correlation between the rankings produced by the two methods.

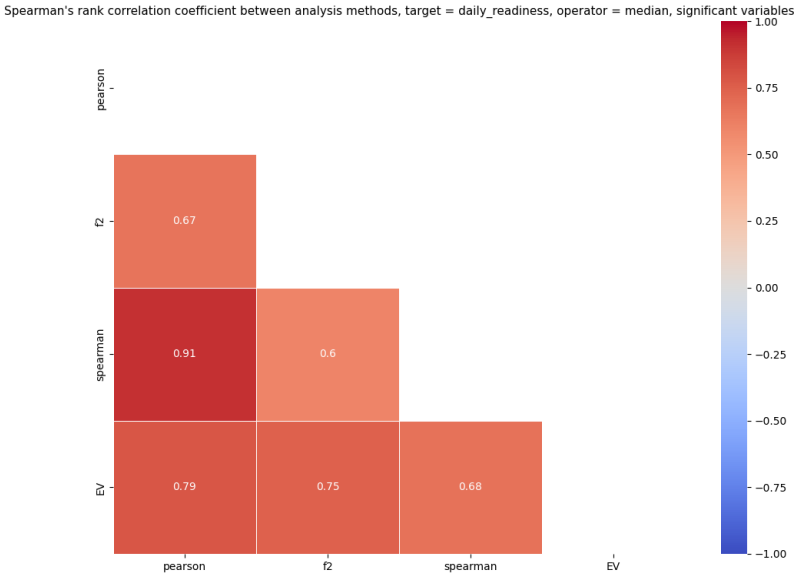


Figure 17: Correlation matrix for analysis methods with daily readiness as target variable

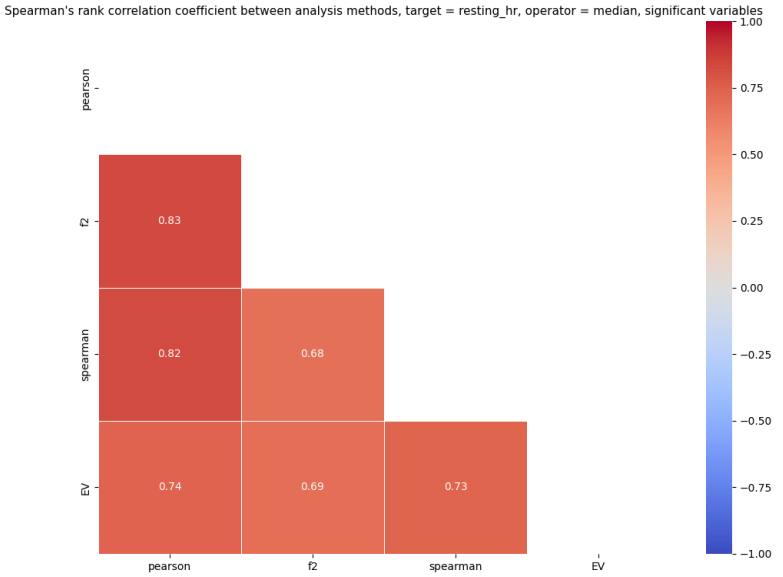


Figure 18: Correlation matrix for analysis methods with resting heart rate as target variable

6 Discussion

In this section, the results will be interpreted. I will also cover the limitations that were encountered during this research.

6.1 Interpretation

The interpretation of the findings from the analysis of the dataset yields several noteworthy observations that contribute to our understanding of the relationships and patterns within the context of speed skating athletes. These key interpretation points cover the nature of the data, the influence of variables, and the efficacy of analytical methods.

Firstly, the prevalence of linear trends in the relationships identified in the data underscores a certain degree of predictability and consistency. The majority of relationships exhibit linearity, emphasizing the straightforward and direct associations between variables. It is crucial to note that non-linear relationships, when observed, rarely provided better results than the results gather by using a linear model. While statistically significant, the results for non-linear models such as quadratic fitting evoke questions about the practical significance of these non-linear trends. This suggests that, within the confines of this dataset, the linear model is a reliable and robust representation of the patterns exhibited by the variables under consideration.

A particularly interesting insight emerges when examining athletes within the same speed skating discipline. The results suggest that there is a heightened similarity in the way athletes within a specific discipline react to various variables. However, caution is warranted when drawing conclusions from this observation due to the limitations imposed by the number of athletes and the temporal span of the dataset. These limitations will be discussed in more detail in the next section. The nuanced nature of athlete responses may be better understood with a more extensive dataset, ensuring a representative and diverse sample.

Furthermore, the ranking of variables generated through the Pearson correlation coefficient aligns significantly with rankings obtained through other methods. The correlation between methods suggests a high degree of agreement across different analytical approaches, reinforcing the robustness and reliability of the identified relationships. The consistency in variable rankings provides confidence in the validity of the findings and reinforces the notion that the Pearson correlation coefficient serves as a dependable tool for assessing the strength and direction of associations in this particular context.

When looking at aggregation of data points and time shifting of the data, the results do not appear to show a clear preference for either method. For the linear approach, 60 aggregated variables showed a significant relationship and 61 time shifted variables showed a significant relationship. Consequently, based on the observed results, no clear preference emerges between the methodologies of data aggregation and time shifting with regard to the establishment of significant relationships within the linear context.

In the examination of methodological comparisons with daily readiness to train as the target variable, the Pearson correlation coefficient exhibits a notable positive correlation with the explained

variance (Subgroup Discovery) and Spearman's rank correlation coefficient, reflecting values of 0.79 and 0.91, respectively. Furthermore, a moderate positive correlation of 0.67 is observed between Pearson correlation and quadratic fitting (f2). These correlations may substantiate the utilization of the Pearson correlation coefficient in conjunction with the non-linear technique of quadratic fitting.

In the analysis involving resting heart rate as the target variable, the Pearson correlation coefficient demonstrates a robust correlation with all three alternative methods. Specifically, high positive correlations are observed with explained variance (Subgroup Discovery), Spearman's rank correlation coefficient, and quadratic fitting, registering respective values of 0.74, 0.82, and 0.83. These findings underscore the efficacy of the Pearson correlation coefficient as a primary analytical tool, capturing a substantial proportion of the outcomes derived from the combined application of all methods and thereby limiting the necessity for supplementary analytical approaches. The selection of a linear metric in conjunction with explained variance emerges as a plausible approach when opting for a two-method analysis.

6.2 Limitations

Throughout this investigation, several limitations were encountered, warranting discussion alongside suggestions for potential avenues of improvement.

Primarily, the quality of the data in this study was notably dependent on the consistency with which athletes completed the questionnaires. Disparities emerged, with some athletes diligently completing all questionnaires while others frequently omitted required entries, resulting in a less reliable dataset with missing values. Notably, the consistency in filling out training logs exhibited substantial variation among athletes. To mitigate this issue, a potential solution could involve mandating athletes to promptly complete the training log after each session, perhaps integrating this task seamlessly into the training sessions by requiring the athletes to fill in the training log at the conclusion of a training session.

Moreover, the temporal scope of this research encompassed a period without speed skating matches, thereby precluding access to match-related data. The inclusion of such data could considerably enhance the study's significance, as race times from matches offer an objective assessment of athlete performance. Conversely, a substantial proportion of the assessed variables relied on subjective evaluations by athletes, introducing potential biases and variability into the dataset. Athlete-provided ratings may be influenced by personal perceptions or emotional states, introducing inaccuracies and inconsistencies that complicate precise conclusions and accurate comparisons. Given that the ultimate goal of the speed-skating team is to optimize race performance, the absence of race time monitoring poses challenges in providing well-founded recommendations for optimizing the training plan.

Additionally, the study was constrained by a limited dataset due to the relatively brief duration of the research period. Comparable studies typically extend over at least one year to derive reliable conclusions. For instance, a similar study analyzing junior speed skaters spanned almost three years [WSV+17]. The time constraints inherent in this research precluded such an extended duration, yet replicating the study over a more prolonged period could yield valuable insights.

7 Conclusions and Further Research

In conclusion, this study undertook a comprehensive examination of the processes involved in the collection, processing, and analysis of a dataset pertaining to speed skating. Subsequent to employing various analytical methodologies on the dataset, the corresponding outcomes for each method were presented. A critical evaluation of the efficacy of these methods was then conducted.

Upon interpreting the results, an attempt was made to address Research Question 1 (**RQ1**). While acknowledging the absence of an unequivocal answer to this question, the prevalent trend observed within the data suggests that the majority of the data within this setting demonstrates a linear association. Non-linear models, such as subgroup discovery, Spearman's rank correlation coefficient, and quadratic fitting, failed to identify substantial non-linear relationships that could offer practical insights. Moving forward, Research Question 2 (**RQ2**) was addressed. Based on the available data and the outcomes of this investigation, linear modeling emerges as the most straightforward and effective approach for analyzing sports training data. The marginal contributions of these additional approaches underscore the efficacy of the Pearson correlation coefficient as a primary analytical tool for this specific set of data. This implies that the linear relationships captured by the Pearson correlation coefficient adequately represent the underlying dynamics, minimizing the need for more complex modeling techniques in this particular context. When opting for a two-method analysis, the use of a linear metric such as the Pearson correlation coefficient alongside with a non-linear method such as explained variance or quadratic fitting seems to be the most effective approach.

These insights contribute to a nuanced understanding of the relationships among variables in the context of speed skating athletes, providing a foundation for future research and exploration. Nonetheless, it is imperative to underscore the need for further validation of these results. The limitations delineated in the preceding section underscore the necessity for replicating this study with a larger sample size and an extended research duration, as these factors may yield divergent results. Additionally, an avenue for future research involves the incorporation of match results data. This expansion could introduce objective performance metrics, offering potential benefits in terms of reliability, validity, and consistency in the analysis of athlete performance.

References

[Ako18] Haldun Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.

[Atz15] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[BL09] Jill Borresen and Michael Ian Lambert. The quantification of training load, the training response and the effect on performance. *Sports medicine*, 39:779–795, 2009.

[BN74] Gunnar AV Borg and Bruce J Noble. Perceived exertion. *Exercise and sport sciences reviews*, 2(1):131–154, 1974.

[DHMVY+22] Ruud JR Den Hartigh, L Rens A Meerhoff, Nico W Van Yperen, Niklas D Neumann, Jur J Brauers, Wouter GP Frencken, Ando Emerencia, Yannick Hill, Sebastiaan Platvoet, Martin Atzmueller, et al. Resilience in sports: a multidisciplinary, dynamic, and personalized perspective. *International Review of Sport and Exercise Psychology*, pages 1–23, 2022.

[FDH+96] Carl Foster, Erin Daines, Lisa Hector, Ann C Snyder, and Ralph Welsh. Athletic performance in relation to training load. *Wisconsin medical journal*, 95(6):370–374, 1996.

[GMMT07] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14–es, 2007.

[KHA18] Deborah Kerr, Patria Hume, and Timothy Ackland. *Recommendations for Conducting Research on Athletes (Large-Scale Survey Case Studies)*, pages 191–204. 01 2018.

[KOH+17] Arno Knobbe, Jac Orie, Nico Hofman, Benjamin Van der Burgh, and Ricardo Cachucho. Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, 31(6):1872–1902, 2017.

[KWC21] W Larry Kenney, Jack H Wilmore, and David L Costill. *Physiology of sport and exercise*. Human kinetics, 2021.

[Mad07] Henrik Madsen. *Time series analysis*. CRC Press, 2007.

[MJW+21] Stephen Mellalieu, Christopher Jones, Christopher Wagstaff, Simon Kemp, and Matthew J Cross. Measuring psychological load in sport. *International Journal of Sports Medicine*, 42(09):782–788, 2021.

[MK21] Marvin Meeng and Arno Knobbe. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*, 35(1):158–212, 2021.

[MKK10] William D McArdle, Frank I Katch, and Victor L Katch. *Exercise physiology: nutrition, energy, and human performance*. Lippincott Williams & Wilkins, 2010.

[Muk12] Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71, 2012.

[PM18] Antonios Pantazopoulos and Manolis Maragoudakis. Sports & nutrition data science using gradient boosting machines. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–7, 2018.

[WSV+17] Rikstje Wiersma, Inge K Stoter, Chris Visscher, Florentina J Hettinga, and Marije T Elferink-Gemser. Development of 1500-m pacing behavior in junior speed skaters: a longitudinal study. *International Journal of Sports Physiology and Performance*, 12(9):1224–1231, 2017.