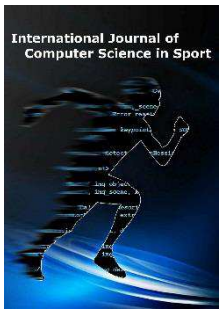
International Journal of Computer Science in Sport

Volume 22, Issue 1, 2023

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2023-0005



Detecting Outliers in Cardiopulmonary Exercise Testing Data of Ski Racers – A Comparison of Methods and their Effect on the Performance of Fatigue Prediction

Baumgartner, N.[1], Kranzinger, C. [1], Kranzinger, S. [1], Snyder, C.[2,3], Stöggl, T. [2,3], Resch, B.[4,5]

[1] Salzburg Research Forschungsgesellschaft mbH, Salzburg, Austria

[2] Red Bull Athlete Performance Center, Thalgau, Austria

[3] Department of Sport and Exercise Science, University of Salzburg, Austria

[4] Department of Geoinformatics – Z_GIS, University of Salzburg, Austria

[5] Center for Geographic Analysis, Harvard University, Cambridge MA, USA

Abstract

In sports science, cardiopulmonary data is used to assess exercise intensity, performance and health status of athletes and derive relevant target values. However, sensors may produce flawed data and data may include a wide variety of artifacts, which could potentially lead to false conclusions. Thus, appropriate and customized pre-processing algorithms are a vital prerequisite for producing reliable and valid analysis results. To find adequate outlier detection methods for this type of data, we compared three algorithms by applying them on seven ergospirometric measures of junior ski racing athletes and applied a model to predict fatigue during skiing based on the pre-processed data. While values that lie outside a realistic spectrum were consistently labelled as outliers by all methods, and mean values and standard deviations changed in similar ways, methods differed from each other when it comes to changing trends, recurring patterns, and subsequent outliers. Decomposing the sensor data into different components (trend, seasonality, remainder) before dealing with outliers increased average predictive performance the most. However, pre-processing remarkably improved prediction results for certain study participants and not for others. Thus, handling outliers correctly prior to deriving information from ergospirometric data is recommended but more research should be conducted to find methods that achieve more consistent improvement.

KEYWORDS: CARDIOPULMONARY EXERCISE TESTING, ERGOSPIROMETRY, OUTLIER DETECTION, SKIING, FATIGUE PREDICTION

Introduction

Data obtained from cardiopulmonary exercise testing, also called ergospirometric or spiroergometric data, provide a basis for various applications in medicine and sports science. For instance, Broich, Sperlich, Buitrago, Mathes and Mester (2012) argue that results of ergospirometric tests have great value for optimization of football players' performance by designing and evaluating individual training programs and Epishev, Korableva, Alferova, Yakhin and Episehva (2019) forecasted further performance enhancement of skiers through ergospirometric data.

However, these sensor-generated data often contain values influenced by various artifacts, such as motion or electrical artifacts, which could potentially result in misleading conclusions if not handled appropriately. According to Robergs, Dwyer and Astorino (2010), the typical approach to data pre-processing in sports science is taking mean values of moving time or breath windows, with typical sizes of 15 to 60 seconds or five to 21 breaths, respectively. Averages are often computed as the arithmetic mean, which is, however, considerably influenced by potential outliers in the data.

Errors in spirometry, in which the patient rapidly exhales, and one breath is analyzed based on flow-volume loops and volume-time curves, have been studied by Luo, Whitmire, Stout, Marteson and Patel (2017). Regarding ergospirometry however, only few studies described how they cleaned outliers when using therefrom derived measures. In a study aiming to evaluate test-retest reliability of certain cardiopulmonary parameters, Barroso et al. (2019) labelled values above $\mu + k * \sigma$ and below $\mu - k * \sigma$ as measurement errors (where μ and σ correspond to the arithmetic mean and standard deviation of the entire trial, respectively, and $k = 1$ is a pre-defined factor). Mapelli et al. (2022) used a so-called interquartile range criterion, excluding observations above $q0.75 + k * IQR$ and below $q0.25 - k * IQR$ (where $q0.25$, $q0.75$ and IQR correspond to first quartile, third quartile and interquartile range of the entire trial, respectively, and $k = 2$ is a pre-defined factor) from their analysis.

A more generic approach to finding accurate outlier detection techniques, where different outlier detection methods were compared on two artificially generated data sets representing biomedical data, was presented by Chromińsky and Tkacz (2010). They inserted outliers into the two test data sets by changing 10% of the 80 values in two different ways, thus producing a set with extreme anomalies and another one with milder ones. Grubbs' Test (Grubbs, 1950), Dixon's Test (Dixon, 1950), Hampel's Test and Quartile Method (which corresponds to interquartile range criterion mentioned in the previous paragraph) were applied with different sensitivity thresholds and compared in terms of correctly detected points and computational time. The best results concerning these two factors were yielded by applying Hampel's Test and Quartile Method with high sensitivity (Chromińsky and Tkacz, 2010).

Reviewing state-of-the-art methods for detecting outliers in time series data in general, Blázquez-García, Conde, Mori and Lozano (2021) stated model-based techniques as the most frequently used approach to detecting point outliers, i.e., a single datum that behaves unusually, in a univariate time series $X = \{x_t\}$. The idea is that a data point x_t, which substantially deviates from it expected value $\hat{x}_t$ should be considered an outlier. The expected value is computed using past, current, and future data via estimation models, which correspond to piecewise constant models in the simplest case but can include more elaborated procedures. Other techniques mentioned are density-based methods, which detect points with less than a certain number of points in a specified neighborhood, and histogramming, a technique that relies on minimizing the error of a histogram representation of the data (Blázquez-García et al., 2021).

This study aims to find suitable outlier detection techniques to be applied to ergospirometric data

prior to further processing tasks, thus supporting correct, valuable predictions via accurate data pre-processing. More precisely, it focuses on data collected in field tests, which is characterized by increased numbers of artifacts when compared to standard laboratory testings. Since there is not much information available on outlier detection in ergospirometric data in this particular context, the purpose of this study is to compare different methods on cardiopulmonary exercise data of ski racers that have been selected based discussed literature. It aims to find out how a statistical method (Rosner, 1983), a simple model-based approach (Basu and Meckesheimer, 2007) and a more advanced model-based approach (Hyndman, 2021) differ from each other. Methods are compared in terms of the points marked as outliers, the impact of outlier removal on descriptive measures and the change of performance of a Random Forest (Breiman, 2001) for assessing perceived fatigue after outliers have been removed. Changes in performance measures of the applied predictive model are used to evaluate the effect of outlier removal according to the respective technique.

Methods

Data collection

Data were collected in a laboratory study on an indoor ski treadmill allowing consecutive turns on an area of 9.5 m × 5.0 m (BIG Maxx ISB 200, MaxxTracks) of the ski middle school Windischgarsten. The study was approved by the local ethical committee (EK-GZ: 11/2018) and conducted in accordance with the Declaration of Helsinki, after having received informed, written consent from all participants. Junior ski-racing athletes, who all were familiar with skiing on this apparatus, were instructed to maintain a turning frequency of 1.2 Hz while breathing through a portable metabolic system (K5, COSMED). Heart rate, breathing patterns and exhalation gas data were collected breath-by-breath. Test trials lasted between 30 and 45 minutes per participant depending on how long the required turning frequency could be maintained.

Additionally, athletes rated their perceived fatigue on a rating-of-fatigue (ROF) scale developed by Micklewright, St Clair Gibson, Gladwell and Al Salman (2017) every two minutes during the trial. They gave a numerical response ranging from 0 to 10, where 0 would mean that the person does not feel fatigued at all and 10 would indicate total fatigue and exhaustion.

Data of three female and six male participants (age: 13.2 ± 0.4; height: 157.8 ± 7.7 cm; weight: 45.9 ± 7.9 kg) were used, after excluding the data of one participant with a fall and one with values for cardiac data that lay outside a realistic physical spectrum.

Choice of parameters

The study used values of seven ergospirometric parameters (see Table 1) for the comparison of outlier detection methods by treating each of them as a univariate time series.

Figure 1 shows violin plots giving information about the distribution of unprocessed, raw data of the parameters per person. The distributions of all athletes are depicted in one plot for each ergospirometric feature showing that there are extreme outliers for RR and RQ data. While values of respiratory variables (RR, TV, VE, VO_2, VCO_2, and RQ) are approximately symmetrically distributed, the distribution of HR data was considerably skewed left across all athletes despite P09. Furthermore, arithmetic means and the variances of the measures differ notably between the study participants. The outlier detection methods, described in the following section, were chosen based on these characteristics of the data.

Table 1. Overview of chosen parameters (COSMED, 2003).

Measure (abbreviation)	**Description**
Respiratory rate (RR)	Breaths per minute (1/min)
Tidal volume (TV)	Air inhaled during one breath (l)
Ventilation (VE)	Air inhaled per minute (l/min)
Oxygen uptake (VO_2)	Oxygen consumed per minute (ml/min)
Carbon dioxide production (VCO_2)	Carbon dioxide produced per minute (ml/min)
Respiratory quotient (RQ)	Ratio of VO_2 to VCO_2
Heart rate (HR)	Heart beats per minute (1/min)

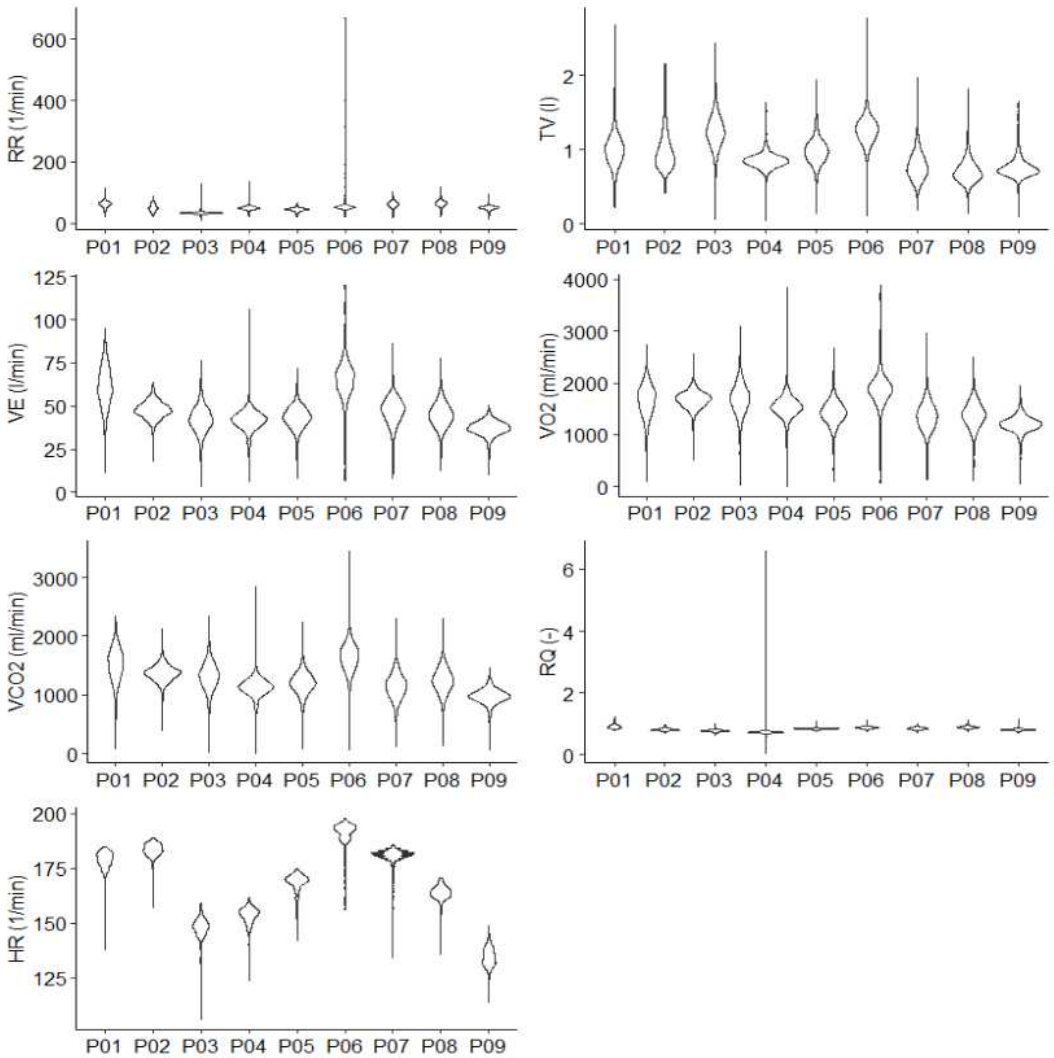


Figure 1: Violin plots of unprocessed ergospirometric parameters by person, RR – Respiratory rate, TV – Tidal volume, VE – Ventilation, VO_2 – Oxygen uptake, VCO2 – Carbon dioxide production, RQ – Respiratory quotient, HR – Heart rate.

Outlier detection methods

The following three outlier detection methods were separately applied to all metrics of every person. For comparison of the detected points, we created plots of the time series with all detected outliers marked in different colors representing the respective method that identified them. For comparison with respect to the influence on descriptive measures and the performance of prediction of fatigue, the marked points were removed from the data set.

The three methods were chosen based on related literature. Since in spirometry, many approaches search for outliers at the lower and upper end without taking the time component into regard, we apply Rosner's test. However, to exploit characteristics of time series data, the other two methods are model-based as these are the most common approaches according to Blázquez-García, Conde, Mori and Lozano (2021). For calculating the expected value, two-sided median method considers a local neighborhood of points, while *tsoutliers()* takes further features of the time series (trend component and seasonality) into account. The choice yields a comparison of three methods, each with a different underlying outlier detection concept.

Rosner's test

Rosner's test (RT) was developed by Rosner (1975) and is designed to detect multiple outliers in a univariate data set. It assumes an underlying normal distribution and detects outliers on the lower and upper end. The used version is based on the generalized extreme Studentized deviate (ESD) (Rosner, 1983). The hyperparameter *k*, which serves as the upper bound of suspected outliers was determined by the number of points outside the interval $[q0.25 - 1.5 * IQR, q0.75 + 1.5 * IQR]$ of the used series of observations.

Two-sided median method

The method proposed by Basu and Meckesheimer (2007) uses the median value of a certain neighborhood of data points to determine outliers. As only the general principle of two-sided median method was introduced in their paper, it was slightly adapted and extended for this application.

For every data point in a series of observations X, a neighborhood of points consisting of the seven previously and seven subsequently recorded data points, was defined. The expected value $\hat{x}_t$ of a respective point x_t was set to the median of this set of points and then compared to the actual observation. If the absolute difference between the actual and the expected value exceeded the threshold $\tau = 6 * MAD$ (where MAD denotes the median absolute deviation of X), x_t is concluded to be an outlier.

tsoutliers() function

The function *tsoutliers()* (TSO) from the R {forecast} package (Hyndman et al., 2022) decomposes the time series X into trend, seasonal and remainder components and detect its outliers based on the remainder series R. The outlier criterion used on R is an interquartile range criterion, where x_t is marked as an outlier if its remainder component r_t lies outside the interval $[q0.25 - 3 * IQR, q0.75 + 3 * IQR]$ of the remainder series R (Hyndman, 2021).

Predictive model

To evaluate the impact of outlier removal on the predictive performance of the dataset as a second part of this study, a fatigue prediction model is applied to the data. In another study using the same data, Kranzinger, Kranzinger, Snyder and Stöggl (2022) explained the rating-of-fatigue (ROF) via ergospirometric measurements by applying a classification model. The used model (Random Forest (Breiman, 2001)) was slightly adjusted to this study's objective to analyze the

influence of outlier removal on the performance of this model and subsequently applied to differently pre-processed data.

The research presented in this paper used seven features (RR, TV, VE, VO_2, VCO_2, RQ, and HR) to train the Random Forest. Tuning of hyperparameters was performed by 10-fold cross-validation. Since the target variable was originally indicated on a scale from 0 to 10 by the study participants, we grouped them into three classes for prediction purposes: mild fatigue (ROF: 0-3), moderate fatigue (ROF: 4-7), and severe fatigue (ROF: 8-10). Figures 2 shows the originally recorded received fatigue ranging from 0 to 10 by study participant and Figure 3 shows the corresponding grouped values.

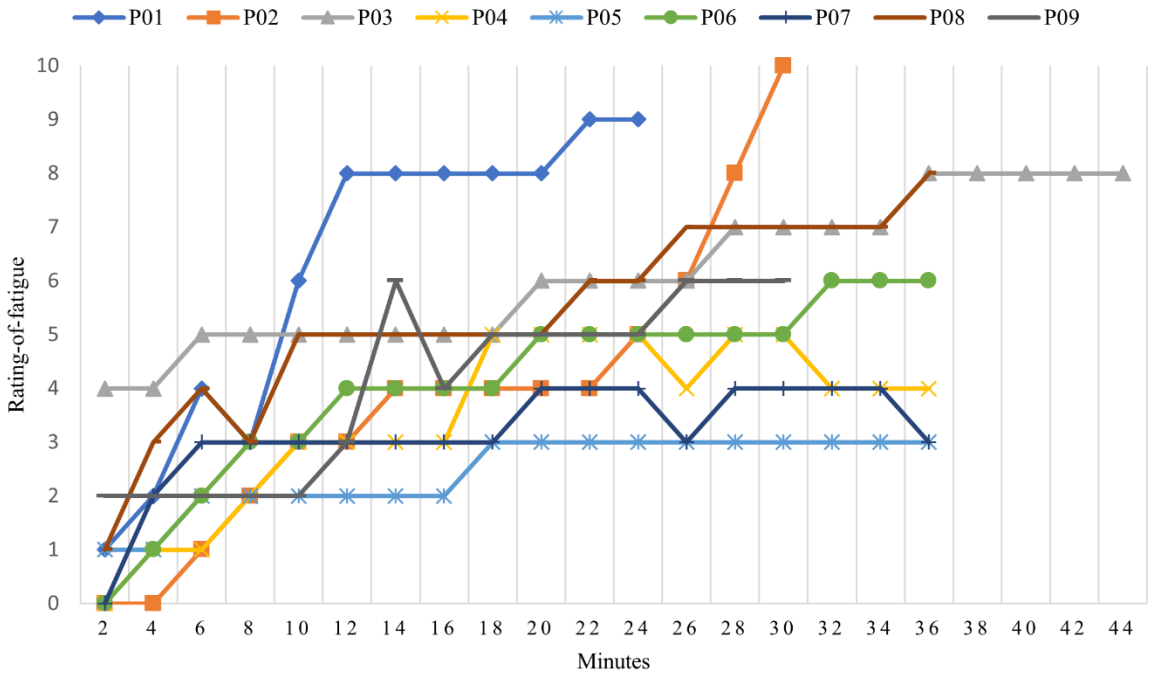


Figure 2. Perceived fatigue of the study participants measured on the rating-of-fatigue scale.

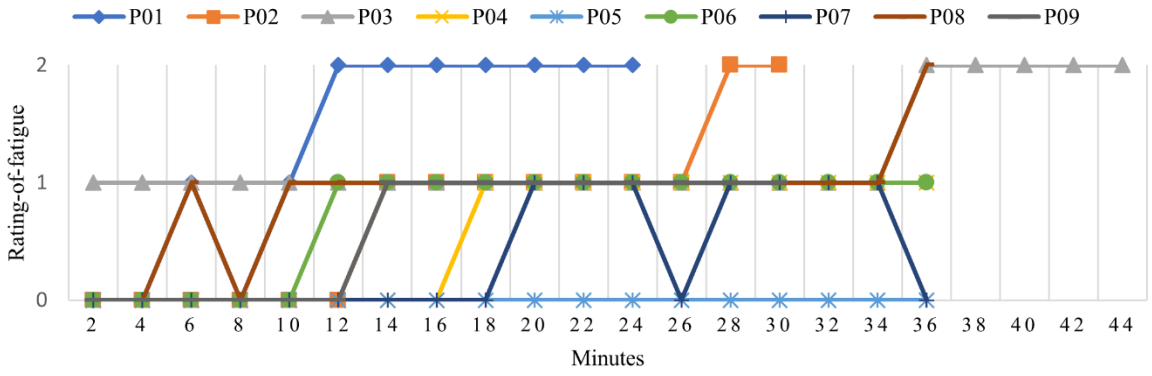


Figure 3. Perceived fatigue of the study participants measured on the rating-of-fatigue scale grouped into three classes, 0 – Mild fatigue, 1 – moderate fatigue, 2 – heavy fatigue.

Explanatory features for deriving the class were computed by taking the arithmetic mean of the respective variable for the timeframe of 30 seconds before the perceived fatigue was assessed.

To investigate the effect of outlier removal on predictive performance, the model was trained

and tested on four different input data sets: on the one hand on the raw sensor data and on the other hand on the three pre-processed time series, where outliers had been previously removed according to the three discussed techniques. We determined the performance of our model via leave-one-subject-out cross validation, resulting in nine test data sets (one for each study participant) and compared it in terms of mean accuracy values, mean Cohen's Kappa values and their respective standard deviations.

Results

Detected points

Of the total 122,394 data points (across all variables of all test persons), RT found 863 outliers, while TSM and TSO identified 775 and 798 points as outliers, respectively. Of these determined points, 376 were labelled as outliers by all three methods. RT and TSM found 392 points in common, RT and TSO 451, and TSM and TSO 579. Thus, according to these numbers, the most similar methods were TSM and TSO.

Table 2 and Table 3 present the number of detected outliers across all measurements by person and by parameter, respectively. Results are compared between the three chosen algorithms with the last row indicating how many points have been marked as outliers by all methods.

Table 2. Number of detected outliers by person: Comparison between different algorithms used for outlier detection, RT – Rosner's test, TSM – Two-sided median method, TSO – *tsoutliers()* function.

Method	**P01**	**P02**	**P03**	**P04**	**P05**	**P06**	**P07**	**P08**	**P09**
RT	46	54	108	88	141	90	222	55	59
TSM	39	36	113	97	82	159	39	72	138
TSO	44	44	130	74	60	211	80	63	92
Common	17	14	72	40	44	65	34	40	50

Table 3. Number of detected outliers by parameter: Comparison between different algorithms used for outlier detection, RT – Rosner's test, TSM – Two-sided median method, TSO – *tsoutliers()* function, RR – Respiratory rate, TV – Tidal volume, VE – Ventilation, VO_2 – Oxygen uptake, VCO_2 – Carbon dioxide production, RQ – Respiratory quotient, HR – Heart rate.

Method	**RR**	**TV**	**VE**	**VO_2**	**VCO_2**	**RQ**	**HR**
RT	269	169	33	55	53	59	225
TSM	234	239	70	105	75	39	13
TSO	179	232	63	83	65	47	129
Common	78	158	25	45	41	16	13

The most coincident outliers were found in the data of P03, while the biggest difference between the used algorithms, as measured as the ratio between the number of the least and most detected points) was observed in P07. Thus, data of these two participants are used to depict exemplary outlier detections in the following passage. Furthermore, major differences could be observed between the methods for HR data, where RT marked 225 points as outliers and TSM detected only 13.

Figures 4 and 5 illustrate detected outliers in all parameters for P03 and P07, respectively.

In the following, findings [F] emerging from the exemplary visualizations of the two chosen study participants are presented and numbered to later be discussed and referenced as F1 to F6 in the discussion section.

[F1] Extreme outliers (points that heavily deviate from most other observations, as seen in RR of P03 (Figure 2)) were consistently detected by all three used methods.

[F2] The points during drop in HR data of P03 after roughly 400 seconds were only labelled as outliers by TSO. The other two detection algorithms regarded those two observations as regularities in the time series.

[F3] Especially when the overall trend of one feature changed, RT often was the only technique that marked certain points as outliers. This can be observed in HR data of P03 (Figure 2) and in the metrics VE, VO_2, and VCO_2 of P07 (Figure 3).

[F4] RT behaved noticeably differently from TSM and TSO for RR data of P07 (Figure 3). It detected 102 points, whereas the other two methods labelled hardly any points throughout the entire recording of RR as outliers (TSM: 1, TSO: 4).

[F5] A point, which is located relatively close to the global arithmetic mean of the respective observations of VO_2 of P03 (Figure 2), was marked as an outlier by TSM.

[F6] Points marked as outliers in different cardiopulmonary measures of the same trial of one athlete sometimes happen at the exact same point in time (breath). This phenomenon can be observed for VE, VO_2, and VCO_2 data of P07 (Figure 3), where two consecutive breaths, taking place at 1,195 and 1,196 seconds, feature anomalous values for the three mentioned metrics.

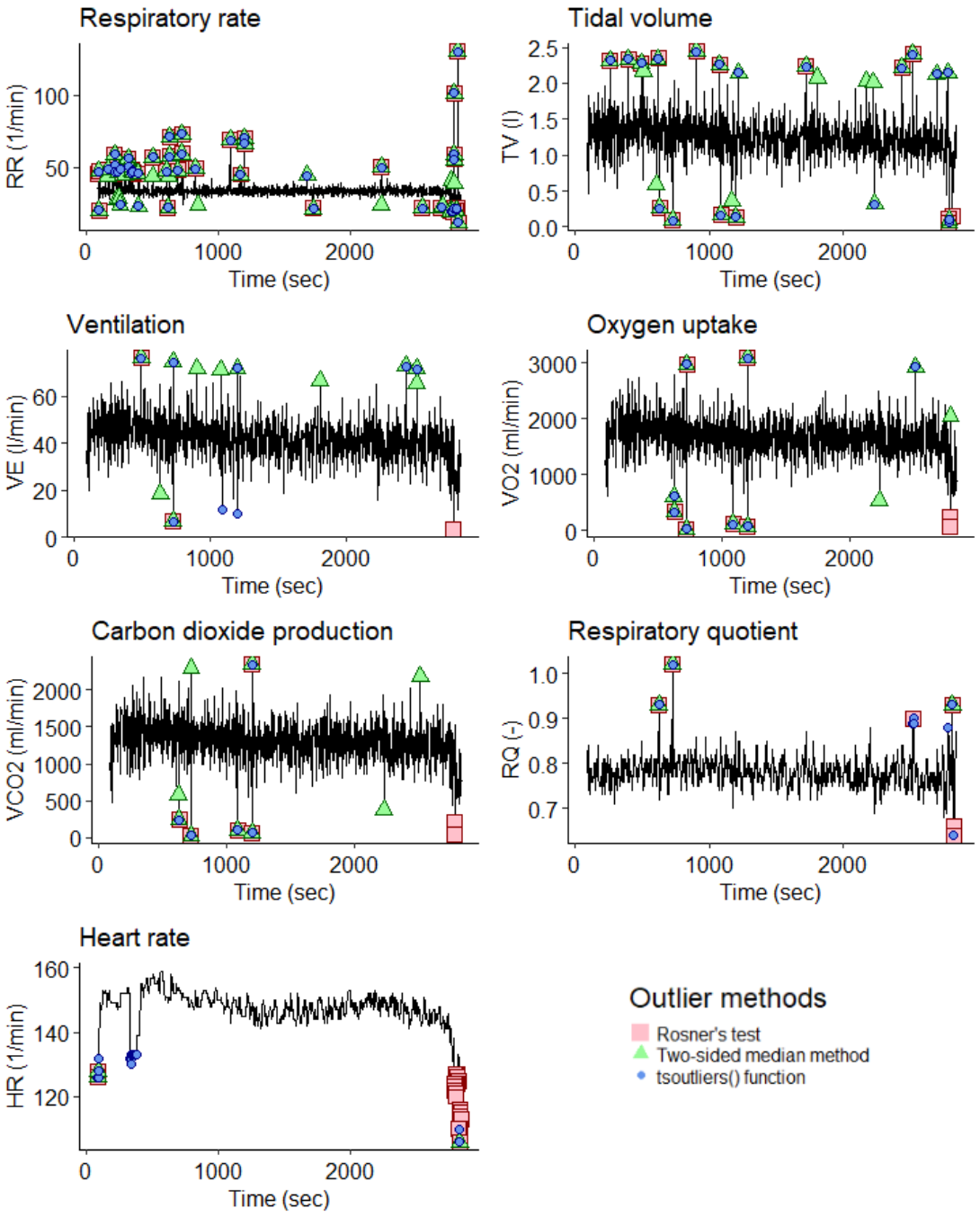


Figure 4. Visualization of detected outliers in P03: Comparison between methods, RR – Respiratory rate, TV – Tidal volume, VE – Ventilation, VO_2 – Oxygen uptake, VCO_2 – Carbon dioxide production, RQ – Respiratory quotient, HR – Heart rate.

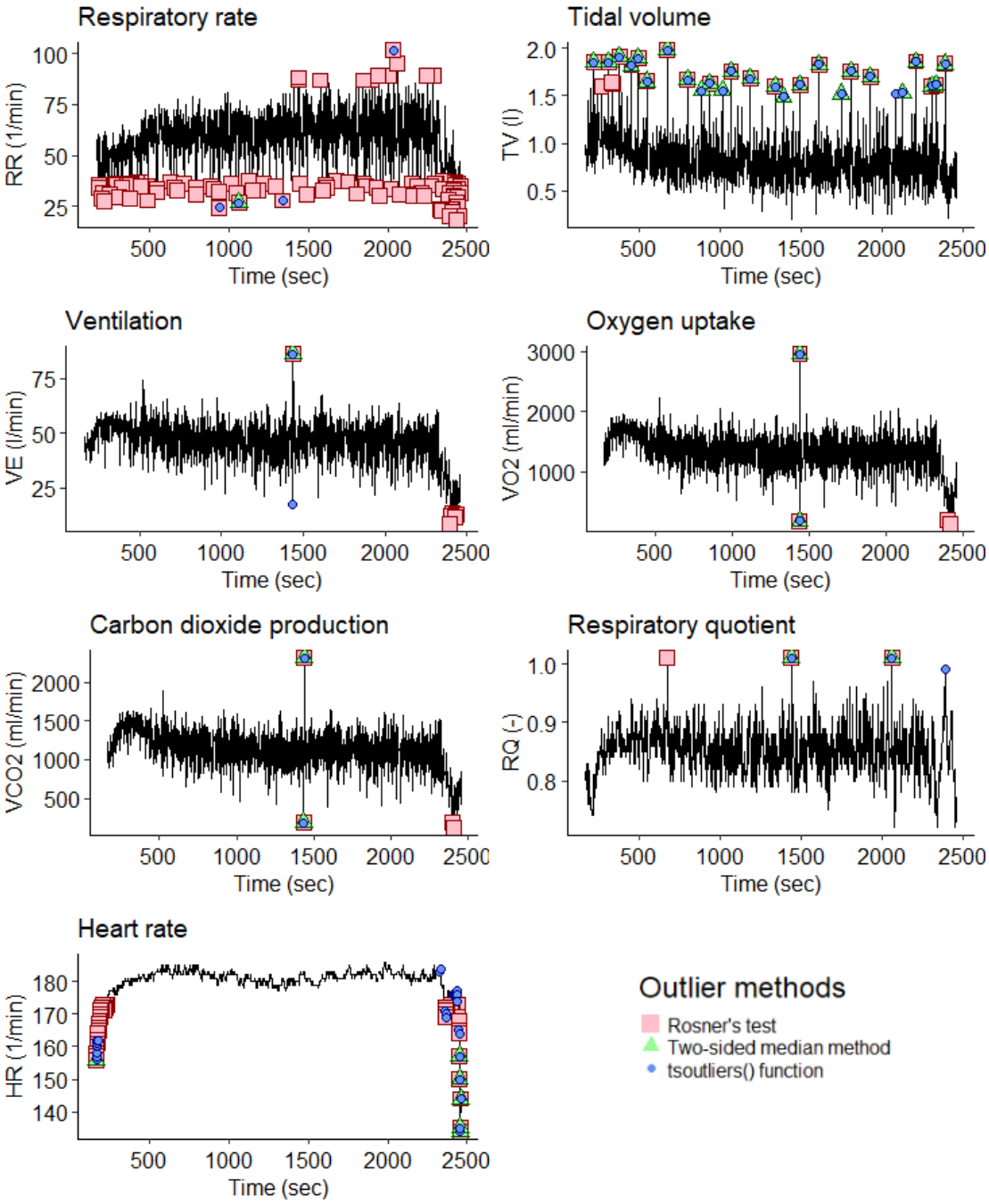


Figure 5. Visualization of detected outliers in P07: Comparison between methods, RR – Respiratory rate, TV – Tidal volume, VE – Ventilation, VO_2 – Oxygen uptake, VCO_2 – Carbon dioxide production, RQ – Respiratory quotient, HR – Heart rate.

Changes in descriptive measures

Outlier removal caused mean values to shift. Although the change differed in terms of intensity, the value tended to shift in the same direction for all algorithms (holds for six out of the seven observed variables). This suggests that the distribution of measurement errors is skewed towards a certain direction. For every parameter, standard deviation decreased after removing the outliers. Minimum and maximum values of the raw sensor data were removed by all three methods in all respiratory measures, but maximum values remained the same in HR data. That is to say that the most extreme values across all study participants of respiratory measures were

labelled as outliers by all algorithms.

Performance of predictive model

Evaluation through fatigue prediction based on the Random Forest scored diverse results for different outlier detection techniques as indicated in Table 4, which shows the mean and standard deviation of the accuracy and Cohen's Kappa across all nine folds of the cross validation.

Table 4. Performance of Random Forest for predicting the rating-of-fatigue: Comparison between different methods used for outlier removal, RT – Rosner's test, TSM – Two-sided median method, TSO – *tsoutliers()* function, CK – Cohen's Kappa, SD – Standard deviation.

Method	**Mean Accuracy**	**SD Accuracy**	**Mean CK**	**SD CK**
None	0.37	0.28	0.08	0.14
RT	0.45	0.28	0.10	0.12
TSM	0.42	0.33	0.11	0.24
TSO	0.45	0.30	0.16	0.27

Mean accuracies ranged from 36.7 to 45.3% and mean Cohen's Kappa values from 0.08 to 0.16. Applying the model to pre-processed data resulted in higher predictive performance for every used algorithm than just applying it to raw data. Regarding mean accuracy values, outlier removal via RT and TSO produced the best prediction results, however, regarding mean Cohen's Kappa pre-processing by TSO outperformed the other two methods. Regarding standard deviation in accuracy values, outlier removal did not produce striking changes. However, TSO had great impact on the standard deviation of Cohen's Kappa values with 0.14 for raw data and 0.27 for data after outliers had been removed. This implies a greater difference of the performance of fatigue prediction between athletes after removing outliers than using unprocessed sensor data, as shown in Table 5.

Table 5. Mean Cohen's Kappa values for prediction of the rating-of-fatigue: Comparison between different methods used for outlier removal, RT – Rosner's test, TSM – Two-sided median method, TSO – *tsoutliers()* function.

Method	**P01**	**P02**	**P03**	**P04**	**P05**	**P06**	**P07**	**P08**	**P09**
None	0.14	0.20	-0.08	-0.04	0.00	0.32	-0.05	0.22	0.00
RT	0.14	0.30	-0.04	-0.05	0.00	0.15	0.13	0.23	0.00
TSM	0.04	0.41	-0.16	-0.04	0.00	0.56	-0.02	0.21	0.00
TSO	0.14	0.28	-0.16	-0.04	0.00	0.72	0.13	0.38	0.00

While the predictive performance remained approximately the same for certain data sets of athletes, the outlier removal according to certain detection methods influenced Cohen's Kappa values for predicting the ROF both positively and negatively for data from other study participants. The method with performance improvement for the most people was TSO by improving four, not influencing for and deteriorating one out of the nine test data sets.

Discussion

The results show that different ergospirometric parameters were differently characterized. While respiratory variables were roughly normally distributed for most study participants, HR data showed a skewed distribution. This different characterization also influenced the number of detected outliers, where results between the three chosen outlier detection methods varied more for HR data in comparison to the respiratory variables. These variations are rooted in the nature of the applied methods, thus raising the question if different ergospirometric features should also be pre-processed differently.

RT does not take the time component of the sensor data into account and only detects outliers that deviate particularly strong from the global mean of the data set. That is why points, which lie rather near to the center with respect to all values, were never marked as outliers by RT, although they seem suspicious to be imprecise measurements (F2). In contrast, it sometimes detected values as outliers when the trend of the time series changed. These points can be regarded as atypical considering the whole time series but seem reasonable when only looking at a local neighborhood (F3). Therefore, we concluded RT not to be a suitable method for data with changing trends, as seen in HR data for instance. Moreover, RT seems to be too sensitive in some situations (e.g., F4), where periodically recurring values are marked as outliers by this technique exclusively, and not by the others.

TSM and TSO, which were the most alike with respect to detected points with 579 coinciding outliers, both take the time component of data obtained by cardiopulmonary exercise testing into account and therefore consider changing trends in the time series when detecting outliers. Thus, they were not prone to identifying fluctuations as seen in VE, VO_2, and VCO_2 data (F3) as outliers. On the other hand, they were sometimes the only methods to detect points relatively near to the global arithmetic mean (F2 and F5). F2, which likely represents a measurement error, still was detected by TSO only and not by TSM because subsequent point showed similar values and influenced the median of the neighborhood used in TSM. In conclusion, although TSM considers the temporality of time series, its weakness lies in detecting subsequent point outliers, which are likely to occur in ergospirometric data. A larger neighborhood (increased *k*) could possibly solve this problem but would reduce the ability to adapt to changing trend thus producing a trade-off problem. By not only considering the trend component but also taking repeating patterns (seasonality component) of the data into account, TSO seems to be the most appropriate approach, which also reflected in the improvement of predictive performance.

Regarding the performance of the classification model to predict the ROF, differences between the mean accuracy and the corresponding mean Cohen's Kappa value were observed. While the former varied from 42 to 45% for cleaned data, mean Cohen's Kappa ranged from 0.10 to 0.16. These Cohen's Kappa values indicate a slight agreement between the ergospirometric measurements and the subjectively perceived fatigue. An unbalanced class distribution of the target value could have been the underlying reason for the difference between the two performance metrics, which raised the challenge of suitable interpretation of results relating to predictive performance. Since Cohen's Kappa takes potential imbalance in the class distribution into account, more importance should be attributed to these values instead of mean accuracy measures.

Observed major differences between the study participants, which were evident in the number of detected outliers, but most importantly in the predictive performance, pose a limitation to the produced results. While the model for predicting perceived fatigue worked well (up to Cohen's Kappa of 0.72 after removing outliers via *tsoutliers()* function) for certain data sets of participants, it had hardly any predictive value for others. Still, results for these people and the

respective change after removing outliers were included when computing mean values for assessing the performance. If there is only a weak association to perceived fatigue, pre-processing the ergospirometric data can never lead to great predictive results. Future research could therefore consider other models that are more reliable for a bigger part of the population and feature a better predictive performance in general. The outlier detection techniques should furthermore be tested in combination with prevalent pre-processing techniques like, for instance, moving averages (Robergs et al., 2010).

Nonetheless, it must be mentioned that it is sometimes impossible to tell whether certain observations are measurement errors and or just represent physiological variability and thus correct values. This was also the case for the data we worked with and posed another constraint to the study, since it was impossible to say with certainty if a point represented a correct measurement or not (despite values that lied outside the realistic range of human physiological data and were constantly correctly labelled as outliers by all three methods (F1)).

According to F6, outliers are recorded in different metrics at the same point in time. This opens the notion of using multivariate outlier detection techniques for future studies. It could be analyzed whether the occurrence of irregular values in multiple variables at once suggests an erroneous measurement (since multiple sensors may have failed) or, quite the contrary, suggests an event of interest for sports scientists (since it may be unlikely for multiple sensors to fail simultaneously). Multivariate techniques could thus not only help in detecting outliers, but also support the decision whether an exceptional data point describes an event of interest or is unwanted data that should be appropriately cleaned.

Conclusion

Since outlier detection in the field of cardiopulmonary exercise testing has so far not been investigated in depth, the goal of this study was to find methods that detect outliers and can therefore support accurate predictions in sports science applications. Multiple techniques, which are prevalent methods for detecting outliers in time series data in general and seem to meet the requirements for effective anomaly detection in cardiopulmonary exercise data, have been compared by applying them to different features of ergospirometric data of ski racing athletes.

The choice of method influences which points are marked as outliers as the number of detected points differs between the methods. Although extreme, unrealistic values are consistently detected by every method, differences occur especially when data contain changing trends or subsequent outliers. Average predictive performance of an example application for predicting fatigue by training a Random Forest improved for every method used for outlier detection, but most significantly (measured by Cohen's Kappa) after applying the algorithm which measures the trend and recurrent patterns in the data (TSO). Nevertheless, it must be stated that the positive effect only happened with certain data sets of athletes and performance remained the same for others, which suggests that more research related to this aspect should be conducted.

Thus, removing outliers prior to performing predictive tasks should be considered during data pre-processing. Especially for data that contain these increasing or decreasing trends and recurring patterns, which is often seen in cardiopulmonary metrics originating from exercise tests, the technique that decomposes the time series in trend, seasonal and remainder component seems to be the best choice. Combining this aspect with the use of multivariate techniques should be studied in the future to potentially achieve even better results.

Acknowledgements

This research was funded by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry for Digital and Economic Affairs, and the federal state of Salzburg. It was a part of the COMET competence center Digital Motion.

This is an extension to the short version of this paper (Baumgartner et al., 2022), which was presented at the IACSS & ISPAS Joint Conference 2022 and submitted to the Springer Proceedings Series Advances in Intelligent Systems and Computing - 13th World Congress of Performance Analysis of Sport (WCPAS2022) & the 13th International Symposium on Computer Science in Sport (IACSS2022).

References

Barroso, M. T. C., Hoppe, M. W., Boehme, P., Krahn, T., Kiefer, C., Kramer, F., Mondritzki, T., Pirez, P., & Dinh, W. (2019). Test-retest reliability of non-invasive cardiac output measurement during exercise in healthy volunteers in daily clinical routine. *Arquivos Brasileiros de Cardiologia, 113*, 231−239.

Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems, 11*(2), 137−154.

Baumgartner, N., Kranzinger, C., Kranzinger S., Snyder, C., Stöggl, T., & Resch, B. (2022). A Comparison of Methods for Automatic Outlier Detection in Ergospirometric Data and their Effect on the Performance of Predictive Models. *Proceedings of 13th World Congress of Performance Analysis of Sport & the 13th International Symposium on Computer Science in Sport.* Manuscript submitted for publication.

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR), 54*(3), 1−33.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5−32.

Broich, H., Sperlich, B., Buitrago S., Mathes, S., & Mester, J. (2012). Performance assessment in elite football players: field level test versus spiroergometry. *Journal of Human Sport and Exercise, 7*(1), 287−295.

Chromiński, K. & Tkacz, M. (2010). Comparison of outlier detection methods in biomedical data. *Journal of Medical Informatics & Technologies, 16*, 89−94.

COSMED. (2003). *K4 b² User manual* (9th ed.). Rome: COSMED Srl.

Epishev, V., Korableva, J., Alferova, T., Yakhin, D., & Episheva, A. (2019). Functional diagnostics in the comparative assessment of physical performance in ski racers to forecast sports performance development. *Proceedings of the 4th International Conference on Innovations in Sports, Tourism and Instructional Science (ICISTIS 2019)*, 53-55.

Dixon, W. J. (1950). Analysis of Extreme Values. *Annals of Mathematical Statistics 21*(4), 488-506.

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *Annals of Mathematical Statistics 21*(1), 27-58.

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice (2nd ed.)*. OTexts. Retrieved June 22, 2022, from https://otexts.org/fpp2.

Hyndman, R. J. (2021). *Detecting time series outliers*. Retrieved June 22, 2022, from https://robjhyndman.com/hyndsight/tsoutliers.

Hyndman, R. J. (2022). *MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns*. Retrieved June 22, 2022, from https://robjhyndman.com/publications/mstl.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2022). forecast: Forecasting functions for time series and linear models. R package version 8.16. https://pkg.robjhyndman.com/forecast.

Kranzinger, S., Kranzinger, C., Snyder, C., & Stöggl, T. (2022). Predicting the rate of fatigue during skiing on a ski treadmill based on ergospirometric data. [Abstract]. *Accepted at the 27th annual congress of the European College of Sports Science - ECSS Sevilla 2022.*

Luo, A. Z., Whitmire, E., Stout, J. W., Martenson, D., & Patel, S. (2017). Automatic characterization of user errors in spirometry. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4239−4242.

Mapelli, M., Salvioni, E., Paneroni, M., Gugliandolo, P., Bonomi, A., Scalvini, S., Raimondo, R., Sciomer, S., Mattavelli, I., La Rovere, M. T., & Agostoni, P. (2022). Brisk walking can be a maximal effort in heart failure patients: a comparison of cardiopulmonary exercise and 6 min walking test cardiorespiratory data. *ESC heart failure*, *9*(2), 812−821.

Micklewright, D., St Clair Gibson, A., Gladwell, V., & Al Salman, A. (2017). Development and validity of the rating-of-fatigue scale. *Sports Medicine*, *47*(11), 2375−2393.

Robergs, R. A., Dwyer, D., & Astorino, T. (2010). Recommendations for Improved Data Processing from Expired Gas Analysis Indirect Calorimetry. *Sports medicine (Auckland, N.Z.)*, *40*(2), 95−111.

Rosner, B. (1975). On the detection of many outliers. *Technometrics, 17*(2), 221−227.

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics, 25*(2), 165−172.