Module 2. Time series analysis

BARÇA INNOVATION HUB
Universitas

Unit 2.1 Time series analysis

In the previous course, we highlighted the advantages of being able to perform calculations with raw data provided by certain technological data collection devices or tools such as GPS, force platforms and accelerometers. We also referred to the concepts of sampling rate or Hz (number of data per second) that the tool records. This type of data, given its continuous recording structure at time intervals, can be classified into a specific data type group. These are time series - that is, those observations recorded over different time intervals.

Although time series analysis in data science is usually associated with areas such as finance and economics, in this module we intend to describe the particularities of some of the data collected in our professional context that meet some of the characteristics of time series as well as the options available to use and make decisions based on their analysis.

This module will not delve into advanced methods for the analysis of time series - we intend to approach real contexts within our field and to present examples of specific situations where we work with this data.

Time series can be classified depending on the total duration of their recording, that is, we differentiate multiple recording points or data collected in a day or a test (GPS or jump test) from periodic evaluations used to make more medium or long term assessments (physical tests, cardiac response, etc.). Both phenomena comply with the premise of being data got repeatedly over time; the difference lies in the time passed between observations.

Micro Time Series

We have decided to call "micro time series" data that represents an action or set of actions recorded over a short and almost uninterrupted period of time. We would like to emphasize that, in order to consider them time series, they must consist of multiple recordings of an intensity signal over time and with a constant recording frequency.

Considering these characteristics of time series, we can find clear examples in the context of sports performance.

Jumps recorded on a force platform:

- Recorded actions: 1-5 jumps, continuously recorded.

- Recorded intensity signal: strength.
- Sampling rate: usually 1000 Hz (intensity signal recorded per second).

Isometric tests with force gauges:

- Recorded actions: 1-5 contractions, continuously recorded.
- Recorded intensity signal: strength.
- Sampling rate: usually >100 Hz (intensity signal recorded per second).

GPS devices:

- Recorded actions: total duration of the session, continuously recorded.
- Recorded intensity signal: position in two coordinates out of which the rest of the metrics are calculated.
- Sampling rate: usually 1-10 Hz (intensity signal recorded per second).

Accelerometers:

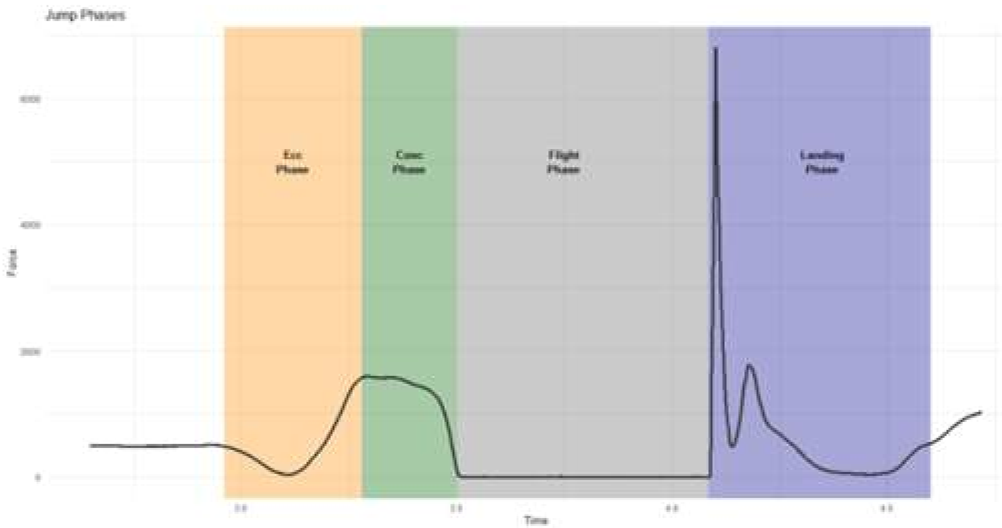
- Recorded actions: total duration of the recording;continuously recorded.

- Recorded intensity signal: position in three coordinates out of which the rest of the metrics are calculated.
- Sampling rate: usually 100 Hz (intensity signal recorded per second).

As we can see, all these examples comply with the characteristics of micro time series. We will not go into further detail, since in the previous course we presented some of the tools used for time series analysis as well as some considerations to ensure the quality of the data we process - the use of RStudio to filter a large volume of data).

Our analysis will aim at data processing and the extraction of characteristics from it. For example, in the graph in Figure 1, from the time series of the force recorded in a jump we use RStudio to determine the different phases of the jump (outlined by colours); Out of this, we can determine characteristics in each of them (total force production, increments, peaks, etc.).

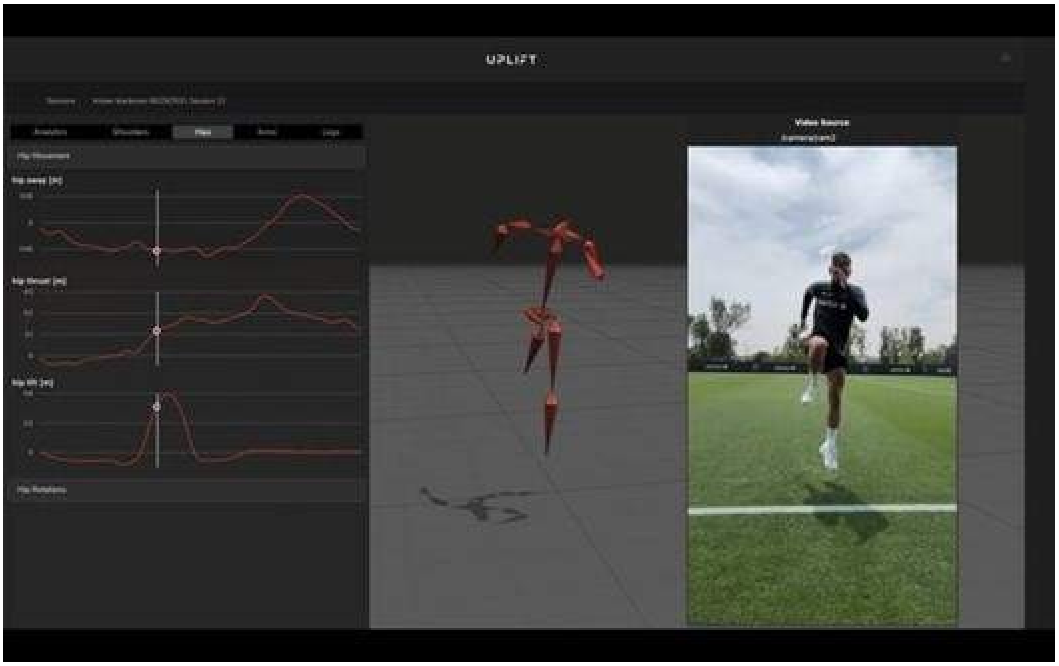
Figure 1: Jump phases



Source: Author's own production

With the development of technology in recent years, examples of this type of time series and their derived analyses continue to grow. Nowadays there is technology capable of recording the position of the different body segments and joints throughout actions or movements, which means that there is no need to use external devices or markers.

Figure 2: Example of a motion analysis tool using video capture



Source: Screenshot taken from Uplift (https://www.uplift.ai/).

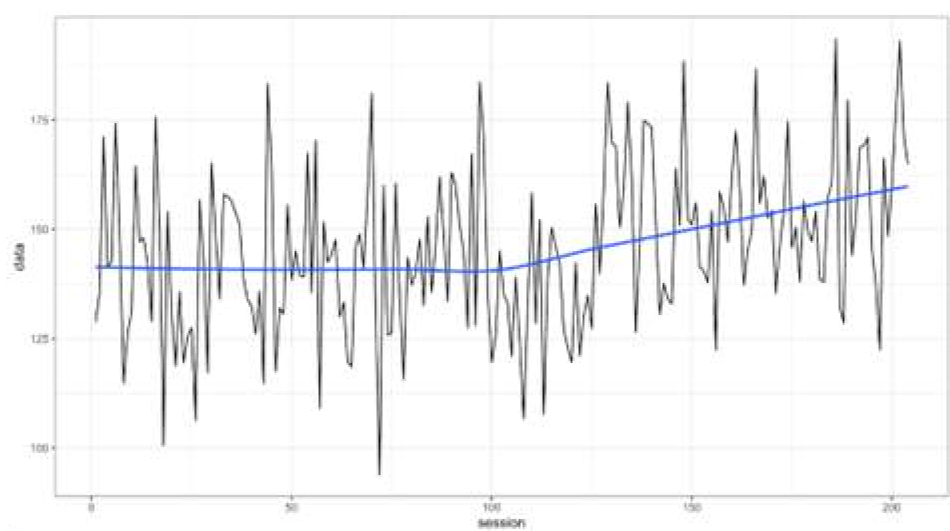
Meso time series

In the initial courses, we highlight the importance of recording the elements related to the question we want to answer in a structured way, establishing regular evaluations to detect changes, or knowing the variability of our players to determine interventions or changes in our periodization or training prescription. In sports where the competitive period is of long duration, we need to know the athlete’s changes over time, since we may continuously record data on a daily (through GPS) or weekly basis (through jump test), etc. We have assigned the term "meso time series" to time series that have been recorded periodically but not continuously, in a period of time that may last between weeks and months.

A common question to answer when we assess this type of data is whether there are changes in the athlete's response, as it may be an indication of the athlete's adaptation to the load and therefore, it may be regarded as an improvement; or of a possible state of fatigue if these values have been reduced or if some event has occurred that caused modifications in the player's response.

Below, there is an example of recording a value throughout different training sessions, recorded on a daily basis. One of the main applications of time series analysis is trend data analysis, i.e., what direction the data follows throughout the records. In this case, this trend is represented by the blue line.

Figure 3: Recording a value across different training sessions

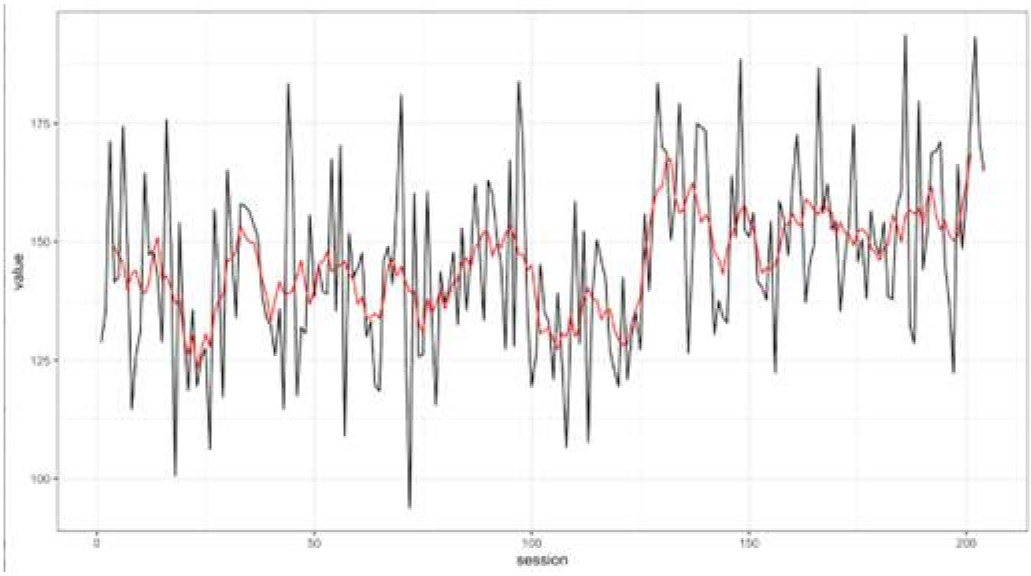


Source: Author's own production

If we were looking for an example applied to the context of sports performance data, these recorded values could be, for example, a variable recorded by a GPS device. This graph could indicate that a player has experienced an increase in their conditional demand over the course of the sessions recorded.

The trend represents the "smoothed" data so that it summarizes what is happening, using data from multiple sessions at once rather than fixating on individual values or spikes. In this way, it is clear that there has been a change towards higher values at some point in the time series (sessions). As we will see in this course, we can use different data smoothing tools such as moving average. These tools aim to summarize what has happened in a time period longer than a single point or session but shorter than the total duration of the record or season. By applying different analysis windows to the moving averages (summary every 3/5/10 sessions, for example), we can see in more detail what is happening in each of those periods with the player's conditional demands.

Figure 4: Drops and peaks in the moving average signal



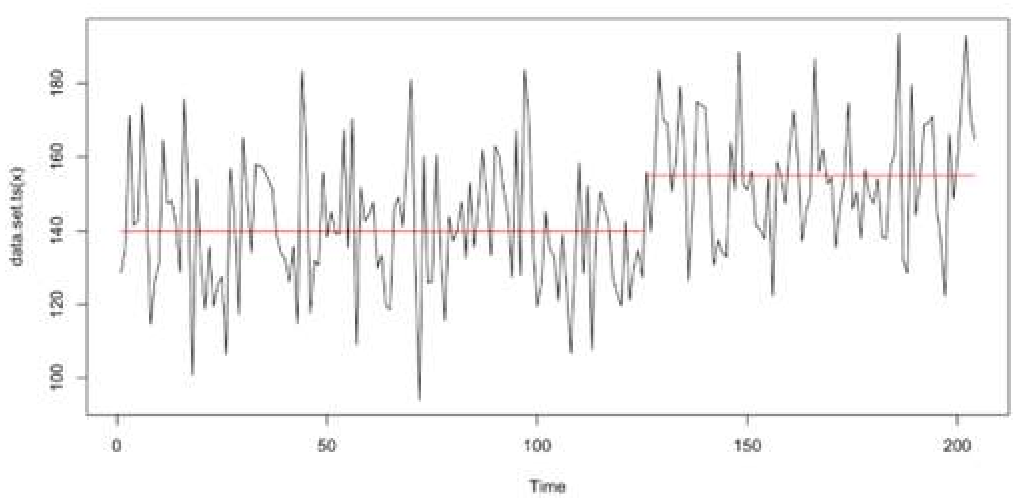
Source: Author's own production

In the chart above, there are distinct drops and peaks in the moving average signal (red line). One way to apply this type of analysis is to use it to determine if these trends or cycles are aligned with the proposed periodization of training and verify that we are performing a load undulation week by week (alternating weeks of greater volume or intensity with weeks of load reduction).

Imagine that the previous analysis does not follow the pattern we had planned, that is, we see that there is some undulation in the values throughout the sessions. However, the first chart with the blue line clearly shows that there is an increase in physical demand throughout the sessions, but this increase was not foreseen in our periodization. In this case, we need to determine the possible cause of this increase.

One of the approaches we can use (making use of tools applied to time series analysis) are change detection algorithms. These are functions that we can use in RStudio to apply directly to our data. The use of these functions makes it possible to determine, firstly, whether there has been any significant change in the values recorded during the time series (session by session) and, secondly, the point at which this has occurred, which is essential in many cases. Below, you can see this function applied to the same time series.

Figure 5: Change detection algorithms



Source: Author's own production

The result shows that a change has taken place in session 129, represented by the two horizontal lines of different values and at a specific point. This information may serve as a starting point to look for the reasons behind this change. Continuing with the example of values recorded by a GPS device that represent the player's conditional demand, one reason could be that the coach has changed the player's position at a certain point in the season (in this example, from session 129 onwards). This change could have led to an increase in the load, given the particularities of the position in which the athlete is now playing.

There are many examples of the application of this type of analysis, and the advantages and information they provide are unremarkable: assessing the point at which the player has greater or lesser asymmetry during the sprint (which may have been related to a blow or physical discomfort), pinpointing the point at which there have been mechanical changes in more analytical movements and investigating their cause, among others.

According to the text, which of the following statements are correct about the "meso time series"?

- [ ] They are recorded continuously on a daily basis, like GPS data.
- [ ] They are recorded periodically but not continuously.

☐ They allow us to know the changes of the player over time in periods from weeks to months.

☐ They are only registered weekly with jump tests.

☐ They refer only to data collected during a short competition.

SUBMIT

Variability as a fundamental factor in the context of sports performance

In the context of sports performance, it is important to be aware that any assessment we make on our players or athletes will have a certain variability associated with the test itself (known as "reliability") and the athletes themselves (Hopkins, 2004), that is, the player will display fluctuations considered "normal" in the assessments we make. These variations are also specific to the metric under analysis. For example, a jump test in which we use Peak Power/BM as a performance measure will show a certain degree of fluctuation test by test for each player or group, while other variables may present different variability. There are metrics that will be more stable (fewer changes session by session) and others that will have greater variability. We see this concept clearly in variables analysed through GPS devices in sports such as football. It should be noted that football is a sport with high complexity, in which the contextual factors (opponent, formation, type of training session, etc.) affect the variability of the metrics recorded.

Clubb et al. (2022) analysed the variability of the different team metrics observed in multiple training tasks and intra-player variability. The objective of this analysis was to determine which metrics display the least variability to use as a reference and thus establish differences between the different tasks proposed.

Figure 6: A violin plot of the within-subject CV for each external training load measure across three different SSG formats.

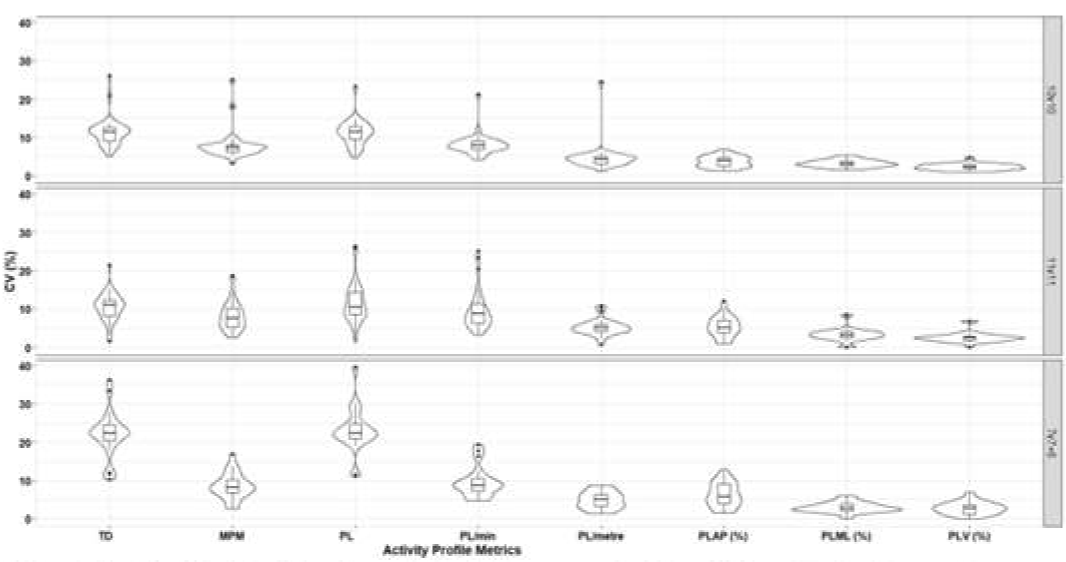


Fig 1. A violin plot of the within-subject CV for each external training load measure across three different SSG formats. TD—Total distance; m/min—metres per minute; PL—PlayerLoad™; PL/min—PlayerLoad™ per minute; PL/metre—PlayerLoad™ per metre; PL_{AP} (%)—% contribution to PlayerLoad™ in the Anterior-Posterior plane; PL_{ML} (%)—% contribution to PlayerLoad™ in the Medial-Lateral plane; PL_{V} (%)—% contribution to PlayerLoad™ in the Vertical plane; CV%—Coefficient of variation.

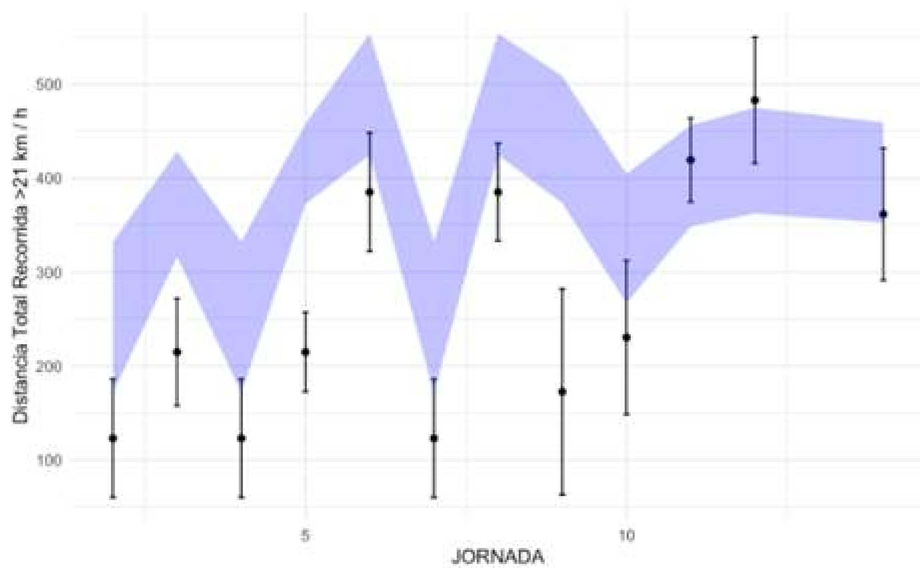
Source: Clubb et al. (2022). https://tinyurl.com/4f6vzn5m

This analysis of variability is very useful when analysing meso time series, as we can use it to determine the moments when there have been changes in the trend of the data we are getting. The example below is a combination of references for the analysis of conditional demand in the metric of distance covered above 21 km/h throughout the days of competition.

The points on the graph represent a player's distance above 21 km/h over the days they have participated in competition, and the vertical line through each point shows the player's variability. The shaded blue area shows the equipment values (average and standard deviation range). In this way, we can evaluate the demand at the group level (blue area) and contribute to the information on whether the match has been more or less demanding. With this reference, we can determine whether the player has been above or below those values. In addition, if the vertical lines are not within that blue area, we can interpret that their values have been well below or above the group demand, since, not even considering their usual variability between matches, the player would still be outside that team range.

We can use this information to make accurate decisions if we determine that this variable is a key performance factor for our team and game plan.

Figure 7: Combination of references for the analysis of conditional demand in the metric of distance covered above 21 km/h throughout the competition days



Source: Author's own production

With this example, we intend to show the differences between analysing the raw data (distance value >21 km/h) day by day and analysing it providing contextual information. If we were to analyse only the raw values, we could consider that match day 7 has been one of the days when the player has had the lowest physical performance, but if we look at what the team has done, we see that it has been a less demanding match at the group level. However, on match day 5 the player had a greater distance covered than on match day 1, and this was much greater than what the team as a whole has done.

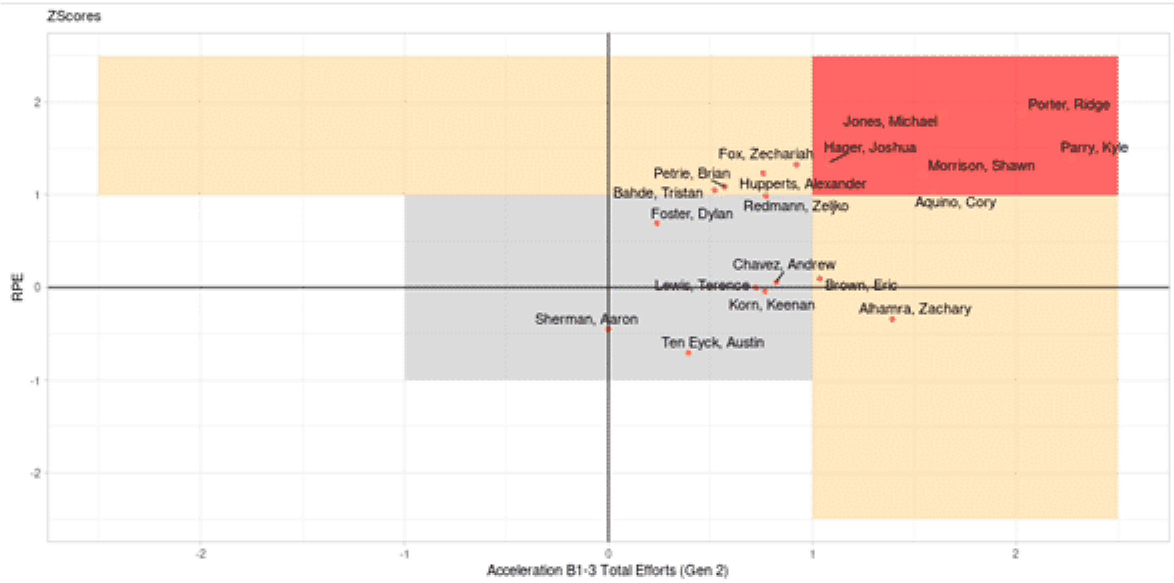
As we know, there are multiple variables used for the analysis when monitoring training load; the same happens with other types of physical performance measures such as tests or kinematic variables. In general, we need to analyse more than one variable at a time in order to have a better representation of what is happening in training or competition. If we add the analysis of variability to each of these metrics, considering the differences between them, the complexity of the comparison increases, since these metrics or variables could have different scales (for example, SPS with a scale of 1 to 10 and total distance with a scale of 200 to 12000 meters) and with different variabilities among them.

In order to standardize these measures and make comparisons between variables and players considering variability, we could use Z-scores. These values measure the position of the raw value (e.g. 200 m distance >21 km/h) from the average and assign a value in standard deviation. Therefore, the values will be in the same units for any metrics analysed with Z-score, usually between -3 and 3. Negative values will indicate a value below the player's usual values, and positive values will indicate higher values. The further away the values are from 0, the more extreme the raw value under scrutiny. As we will see in the video material, RStudio allows for this type of calculation to be carried out directly and efficiently.

Not only can we standardize and make comparisons in the same units for different metrics, but we can also determine that average and that variability. By this we mean that we may not be interested in using the average value of the entire season to determine whether the value of a specific day has been more or less demanding. As players experience changes during the season, based on our context and approach as sport scientists, we could establish that the optimal period as a reference for the most recent data should contain, for example, only the values of the last month.

See below an example applied to the control of the weekly load for a football team. Figure 8 shows two variables: one of internal load (SPS) on the vertical axis, and another of external load (number of accelerations) on the horizontal axis. The scatter plot chart is used to observe which players and in which metric each have got higher or lower values, as well as the relationship between the variables. In this figure we want to represent how demanding the session has been for each of the players if we compare it with the usual values; so that, at the end of the session, we will be able to make decisions based on the results. A Z-score has been applied to each of the values, using the average of each of the players (not the group average) and only in the same type of session (MD-3, i.e. all the sessions carried out 3 days before the competition) as a reference. This approach has been chosen because we know that there are differences between types of sessions depending on the proximity to the match and because each player has specific variability.

Figure 8: Internal Load and External Load Variables



Show 5 entries

Search:

Date	Session	Name	RPE	AlertaY	Acceleration B1-3 Total Efforts (Gen 2)	AlertaX	Alerta	Accion
2021-02-18	MD-3	Fazekas, Michael	NA		70	↑		Valores Bajos
2021-02-18	MD-3	Sherman, Aaron	4.5		71			Valores Bajos
2021-02-18	MD-3	Fox, Zechariah	9.0	↑	76			Revisar Carga Anterior
2021-02-18	MD-3	Brown, Eric	5.5		47	↑		Revisar Wellness de Mañana
2021-02-18	MD-3	Petrie, Brian	10.0	↑	65			Revisar Carga Anterior

Showing 1 to 5 of 21 entries

Previous 1 2 3 4 5 Next

Source: Author's own production

In the graph above we see the players who are placed with high Z-score values: to the right for the external load variable, and on top for the internal load variable. In the table we can see the raw value of each of the variables - this highlights the importance of using Z-scores in a case like the one in the example. The player Eric Brown has made 47 accelerations and is located on the right of the graph for the variable “number of accelerations”; on the other hand, the player Aaron Sherman has had a greater number of accelerations (71), but his value is centred. This indicates that the player Eric Brown, in the MD-3 sessions, usually performs a lower number of accelerations and that the session on that particular day was of greater intensity; however, Brown's Z-score, with a higher number of accelerations, is 0. This could indicate, on the one hand, that the value is close to its average and, on the other hand, that the player is used to bearing this load in this type of sessions or has adapted to it.

Our context and the knowledge of the parameters that may be affecting the variability and changes in the time series data we analyse are essential to draw conclusions and make decisions. In the following example from Altini (2020) (Figure 9), in the analysis of a variable such as HRV (Heart Rate Variability), which has great day-to-day variability, it is essential to establish reference ranges in which these fluctuations are taken into account so as not to make inaccurate decisions. Once this requirement has been met, the rest of the factors that may be affecting the values analysed can be considered.

Figure 9: Analysis of an HRV variable

Source: Altini (2020)

In the context of sports performance, what factors contribute to the variability in the evaluations made to players or athletes?

- [ ] The coach's technique
- [ ] The reliability of the test used
- [ ] The climatic conditions
- [ ] The athlete's normal fluctuations
- [ ] The type of sport practised

SUBMIT

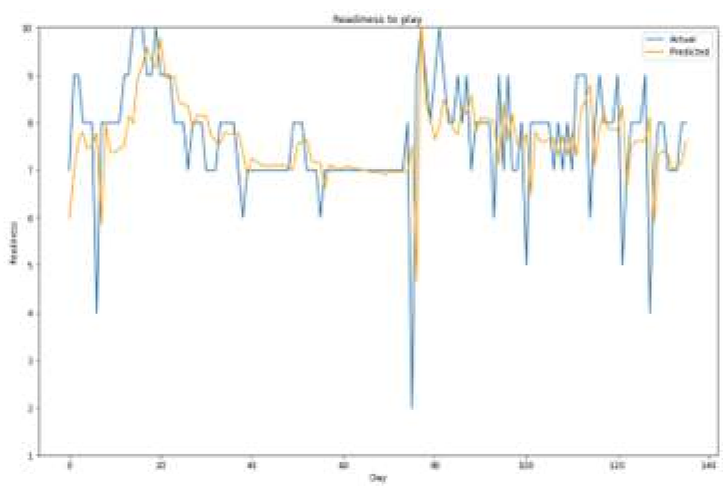
Macro time series

Finally, although we have seen examples where data collection is extensive in time and lasts weeks and months, we want to describe one more category that encompasses those time series with which more complex analyses are carried out or where the volume and the duration of the record is quite significant. This section aims to highlight the importance of establishing long-term data collection plans within the team or organization structure in which we work.

This historical data provides great possibilities for analysis in the future, even with tools that are still in development, but with which we will be able to use that stored data from the past.

We do not intend to deal with advanced algorithms or models for data analysis again. Only the following example (Figure 10) will be shown to highlight that specific analysis methodologies may be applied for time series (in this case, to predict the state of readiness before competition). Although it may be an analysis in a not-so-long time window in which the results are applied (140 days), 2 years of data were used to elaborate the model, and, together with the use of advanced data analysis algorithms, we can consider them within the macro time series analysis.

Figure 10: A player's readiness actual and predicted values



(a)

Source: Kulakou et al. (2022), p. 9

In the following and final case by Mujika et al. (2023), data from the results of different swimming events in official competition in recent years are used to make predictions about the times that will be achieved at the Paris 2024 Olympic Games for each of the positions. This is intended to establish the benchmarks in such competition and it serves as a guide for coaches and professionals who work with swimmers to make adjustments to their programs based on the swimmer's current results. As we can see, this type of information can be crucial for a competition of these dimensions, with the media, economic and sporting implications that each result may have.

Table 1: Example of data used to adjust training programs

Swimming Performance Time Predictions Paris 2024 5

Table 1 Updated Predictions From Both Approaches for the Paris 2024 Olympic Games

	Wu et al[17]		Crowley et al[16]		
Event	**Gold medal**	**Bronze medal**	**1st–3rd**	**4th–8th**	**9th–16th**
Men's events					
50-m freestyle	00:21.09	00:21.43	00:21.38	00:21.63	00:21.87
100-m freestyle	00:47.01	00:47.40	00:47.49	00:47.89	00:48.22
200-m freestyle	01:43.73	01:44.55	01:44.16	01:45.15	01:46.09
400-m freestyle	03:41.25	03:42.35	03:42.97	03:44.70	03:46.61
800-m freestyle	NA	NA	07:40.17	07:46.65	07:49.63
1500-m freestyle	14:34.07	14:35.46	14:36.71	14:51.94	15:00.30
100-m backstroke	00:51.83	00:52.08	00:51.91	00:52.72	00:53.35
200-m backstroke	01:53.32	01:54.72	01:54.64	01:57.34	01:57.99
100-m breaststroke	00:57.40	00:58.23	00:57.75	00:58.81	00:59.49
200-m breaststroke	02:06.17	02:06.83	02:06.64	02:07.12	02:09.28
100-m butterfly	00:49.63	00:50.71	00:50.15	00:50.99	00:51.57
200-m butterfly	01:50.36	01:53.42	01:52.48	01:55.13	01:56.26
200-m individual medley	01:55.03	01:55.85	01:55.79	01:56.77	01:58.54
400-m individual medley	04:06.44	04:08.95	04:08.40	04:11.66	04:14.53
Women's events					
50-m freestyle	00:23.85	00:24.08	00:24.00	00:24.68	00:25.00
100-m freestyle	00:52.01	00:52.32	00:52.28	00:52.74	00:54.14
200-m freestyle	01:53.83	01:54.57	01:53.79	01:56.88	01:57.49
400-m freestyle	03:57.71	04:00.70	03:57.40	04:04.82	04:07.17
800-m freestyle	08:09.86	08:15.40	08:14.58	08:22.62	08:33.21
1500-m freestyle	NA	NA	15:40.16	16:04.48	16:17.26
100-m backstroke	00:57.84	00:58.18	00:58.06	00:59.38	01:00.12
200-m backstroke	02:04.27	02:06.01	02:05.53	02:07.71	02:11.51
100-m breaststroke	01:04.90	01:05.55	01:05.48	01:06.36	01:06.62
200-m breaststroke	02:20.04	02:21.60	02:21.19	02:24.31	02:26.23
100-m butterfly	00:55.51	00:55.98	00:55.49	00:56.44	00:57.85
200-m butterfly	02:04.53	02:05.60	02:05.70	02:07.85	02:10.67
200-m individual medley	02:07.33	02:08.30	02:08.27	02:10.27	02:11.86
400-m individual medley	04:30.41	04:32.88	04:34.99	04:38.39	04:52.44

Abbreviation: NA, not available. In the original work by Wu et al,[17] the men's 800-m freestyle and the women's 1500-m freestyle were not Olympic events. Therefore, no predictions were made for either event.

Source: Mujika et al. (2023), p. 5.

What aspects are important in macro time series analysis?

- [ ] The use of short-lived data
- [ ] The exclusive use of current tools
- [ ] The elimination of historical data

- [ ] The performance of complex analyses

- [ ] The collection of long-term data

SUBMIT

CONTINUE

References

Altini, M. (2020). *The Ultimate Guide to Heart Rate Variability (HRV): Part 2. Medium.* https://medium.com/@altini_marco/the-ultimate-guide-to-heart-rate-variability-hrv-part-2-323a38213fbc

Clubb, J., Towlson, C., & Barrett, S. (2022). *Measurement properties of external training load variables during standardised games in soccer: Implications for training and monitoring strategies*. Plos One, 17(1), e0262274. https://doi.org/10.1371/journal.pone.0262274

Hopkins, W. G. (2004). *How to interpret changes in an athletic performance test*. Sportscience, 8, 1-7. http://www.sportsci.org/jour/04/wghtests.htm

Kulakou, S., Ragab, N., Midoglu, C., Boeker, M., Johansen, D., Riegler, M. A., & Halvorsen, P. (2022). *Exploration of Different Time Series Models for Soccer Athlete Performance Prediction*. Engineering Proceedings, 18(1), 37. https://doi.org/10.3390/engproc2022018037

Mujika, I., Pyne, D. B., Wu, P. P., Ng, K., Crowley, E., & Powell, C. (2023). *Next-Generation Models for Predicting Winning Times in Elite Swimming*

Events: Updated Predictions for the Paris 2024 Olympic Games. International Journal of Sports Physiology and Performance, 18(11), 1269-1274. https://doi.org/10.1123/ijspp.2023-0174

CONTINUE