Module 2. Creating metrics out of raw data

BARÇA INNOVATION HUB
Universitas

Module 2. Creating metrics out of raw data

How is the data we get calculated?

We want to assess the speed of the players in our team. First of all, we need to ask ourselves what aspect of speed we want to assess. Although this may seem simple enough, the test we use will influence the nature of the speed we will assess. We will look at two examples. We may decide to use "timing gates". These devices enable us to time how long an athlete takes from the moment they start the sprint until they cross the finish line. This type of technology usually consists of devices with a laser directed at a reflector in front of them. These two pieces (laser and reflector) form a pair or gate and they are usually placed at the same height - the first pair is placed at the start of the sprint and the second, at the end. The athlete, usually placed slightly behind the first pair of devices, when initiating the movement and crossing through the first gate, will interrupt the signal that the device is emitting against the reflector, which will activate a timer, which will stop the moment the athlete passes through the second gate. The result we will get from this test is a numerical value - the seconds it took the athlete to covel the distance between the two gates.

If we perform this test on the whole team, we will be able to spot which players are the fastest covering, for example, a 40 meters distance, which means they have taken less time to cover the requested distance. However, does the fact that an athlete spent the least time in the test indicate that this athlete has the highest top speed or is really the fastest on the team? The answer is no; a player may have reached a top speed at 20 meters, but may not be able to keep that speed up for the remaining 20 meters. On the other hand, a player who is able to reach a slightly lower top speed and keep it for a longer distance will be able to cover the full distance in less time.

If our objective is to assess our athletes' top speed, then the test used must be different. One option may be to use GPS devices. These devices help measure the athlete's speed in each registered position, as they compute the change of position based on GPS technology. Depending on the type and the brand of the device, the position is measured at different frequencies. The term frequency in these devices refers to the number of frmes per second. That is, if a device records at 1 Hz, that means we have one record of the player's position every second; on the other hand, if a device measures at a 10 Hz frequency, we will have 10 frames recorded per second. Using each of these recorded frames, we can see in which of them the player has achieved a higher speed and thus determine their top speed. In addition, having information about the athlete's position and speed, we could also get the time it took to cover a certain distance, which would result in the same type of analysis as "timing gates".

The aim of this example is not to decide which type of test is best, but to highlight the differences between the measures collected. In the first case we get a single value estimated by simple processing done with a timer while in the second case we record the entire sprint to get information from one of the recorded frames. It is often useful to perform the simplest possible measurements, but also to know the potential that devices with a high frequency of registry have.

Self-assessment: The correct option is highlighted in yellow

Which of the following statements about timing gates are correct?

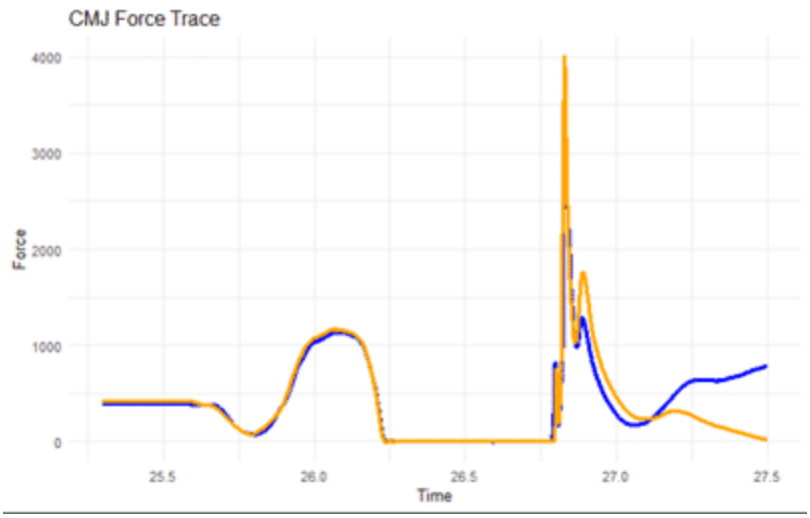
- [ ] A) Timing gates consist of a laser and a reflector that make up a gate.
- [ ] B) The timer is activated when the athlete passes through the second gate.
- [ ] C) The timer stops when the athlete passes through the first door.
- [ ] D) The athlete interrupts the signal when passing

through the first door, thus activating the timer.

SUBMIT

As users of the data provided by different technology companies to measure physical performance variables, we are used to receiving and using "calculated" values from data collected at high frequency. That is, if we perform a jump test measured with a force platform that measures at 1000 Hz, the data that is actually being collected by the platform looks like the following:

Figure 1. Example of data calculation – jump test



Source: prepared by the author

It is difficult to draw conclusions out of this graph at a simple glance; we could see peaks of strength in different parts, but getting information in this way is of little use. It is because of this that different software process this signal to distinguish the different phases of the movement and provide us with information with already calculated variables, such as, the force peak of each phase, the duration, the impulse during the concentric part or the “rate of force development”.

The same goes for GPS devices or accelerometers - the output that we usually get from the different technology companies consists of calculated variables that we will later use in our analysis. It is essential to know where this data comes from and how it has been calculated in order to determine its usefulness in our field.

Unprocessed data as recorded by the device, is also called "raw data". The use of this data has great potential to develop more complex and, above all, tailor-made analyses.

This is an example of the differences between the type of files we get after downloading the raw data (A) and the results calculated by the software (B) for the same test (CMJ jump, 1 repetition). In the raw data file we see few variables/columns, but many observations (more

than 2100). In the case of the calculated variables, we see only one observation (1 jump) but with as many variables as the software allows and we want to use.

Figure 2. Example of data calculation – jump test

A

Time	Left	Right
0	-1.367493	-0.498093
0.001	-1.367493	-0.498093
0.002	0.632507	0.501907
0.003	-1.367493	-0.498093
0.004	0.632507	-0.498093
0.005	-1.367493	-0.498093
0.006	-0.367493	-0.498093
0.007	-0.367493	0.501907
0.008	-0.367493	-0.498093
0.009	-1.367493	-0.498093
0.01	0.632507	-0.498093
0.011	-0.367493	0.501907
0.012	0.632507	-0.498093
...	...	...
+2100 filas	+2100 filas	+2100 filas

B

Athlete	*Test Type*	*Test Date*	*Body Weight [kg]*	*Trial*	*Concentric Mean Power / BM [W/kg]*	*Eccentric Mean Deceleration Force [N]*	*Jump Height (Flight Time) [cm]*
Marco Marquez	CMJ	2/4/2022 00:00	80.94	Trial 1	36.6	1689.5	40.4

Source: prepared by the author

Why use raw data?

Although the efficiency and practicality of using calculated or processed data is clear in the context of daily data analysis, on many occasions we encounter limitations in the software we use. New research often presents new variables that may be useful in our field, potentially using metrics that other brands have, but we are not using. The process of incorporating new metrics by a technology brand can be long and if we are interested in reproducing certain analyses that we consider to have an impact on our team, we must look for a way to carry them out ourselves.

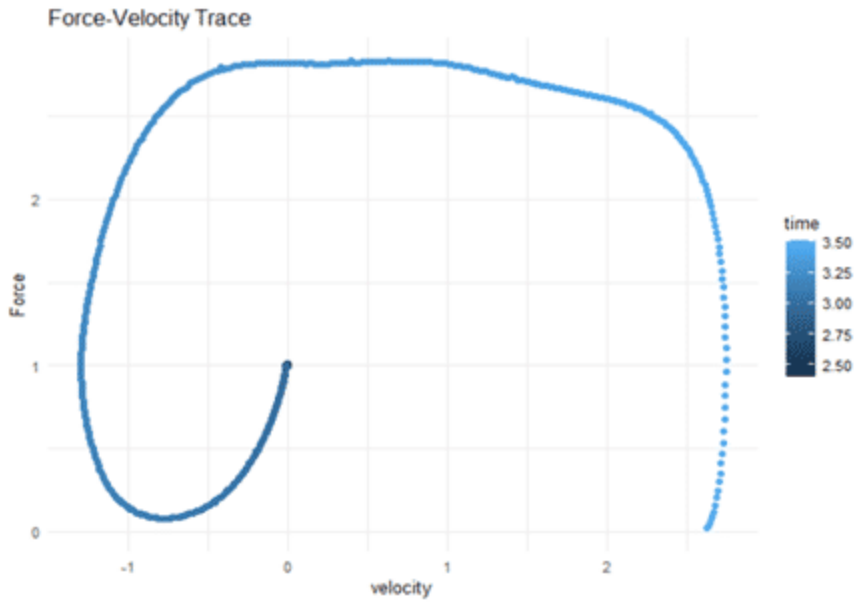
Raw data and knowledge of its processing and analysis will facilitate achieving this objective, will give us greater flexibility in our analysis which will be tailor-made.

Examples of analyses using raw data

Through the video material we learn how to work with raw data, from the specific characteristics of the data import to methods to efficiently analyse them. The examples below show how to use raw data to replicate research and deepen the analysis without having the limitations of the different software.

The CMJ has been a widely used test to assess athletes' changes in performance, fatigue and readaptation processes (Bishop et al., 2023). Calculated variables (values such as RSI, eccentric duration, RFD) are usually used to assess changes, but Gathercole et al. (2015) propose an alternative method to assess fatigue. It uses the raw data of the different jumps to calculate a series of variables derived from the values provided by the force platform and displays the calculated values and how they change throughout the jump. This way, instead of using a single value, we can see how the athlete changes their jumping strategy based on the level of fatigue. If we were interested in using this type of analysis, we should check if the software we use provides this type of visualization. If this is not the case, we can download the raw data, perform the relevant calculations and visualize them on later analysis.

Figure 3. Examples of the use of raw data



Source: prepared by the author

Another example of flexibility on the analysis of raw data may be the use of GPS data. A device that records information throughout the training session or match is a great source of information. In previous modules we have highlighted the usefulness of recording physical performance using these devices and there is extensive literature that justifies their use on the assessment of performance in different sports (**Caro et al.**, 2022).

Self-assessment: The correct option is highlighted in yellow

True or False: The incorporation of new metrics by a technology brand is usually fast, which makes it easier to carry out impactful data analyses on our team.

- ◯ True
- ◯ False

SUBMIT

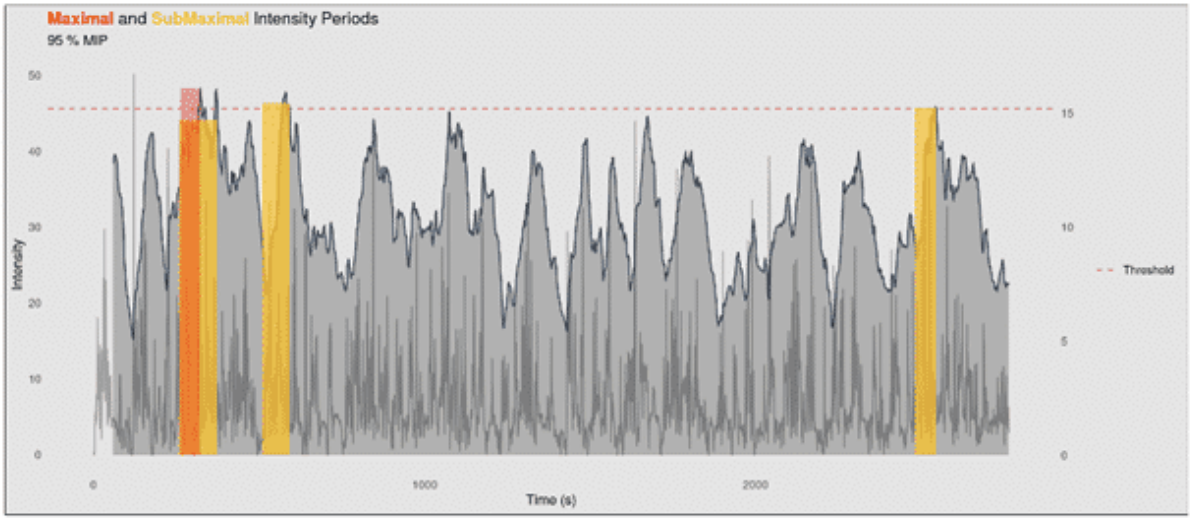
In recent years, different analyses have been developed to assess the phases in which the athlete experiences greater conditional demands during matches (García et al., 2020). These phases are also known as periods of maximum intensity. Knowing the intensity of these phases is of great importance in order to prepare athletes during training, as it has been shown that the average intensity of the match underestimates the specific demands of each position. Based on this approach to performance assessment, complementary analyses, such as submaximum intensity periods (SubMIP), have been developed

(Caro et al., 2022). This approach aims to demonstrate that these phases do not occur in isolation, but occur repeatedly at a similar intensity to those maximum values. In addition, they are position-specific, which justifies their use in workload control during training.

This type of research is relatively new, so many of the GPS data analysis software do not allow for the calculation of these variables. This limitation can be addressed by using raw data from the devices and replicating the methods described in the studies.

In the following graph there is an analysis of the submaximal periods of a football match. This analysis is also present in the video material, but it should be noted that the use of raw data is not exclusively about visualizing a signal or metric (as in the previous example), but that we can get summary values from their analysis and calculations. In the image below we can see the number of submaximal periods, their intensity and their duration. These values may serve as reference in designing tasks that replicate the demands of competition.

Figure 4. Analysis of the submaximal periods of a football match



id	event_duration	intensity	start_time	end_time
1	117.50	14.71	256.70	374.20
2	81.30	15.47	510.40	591.70
3	63.40	15.25	2480.50	2543.90

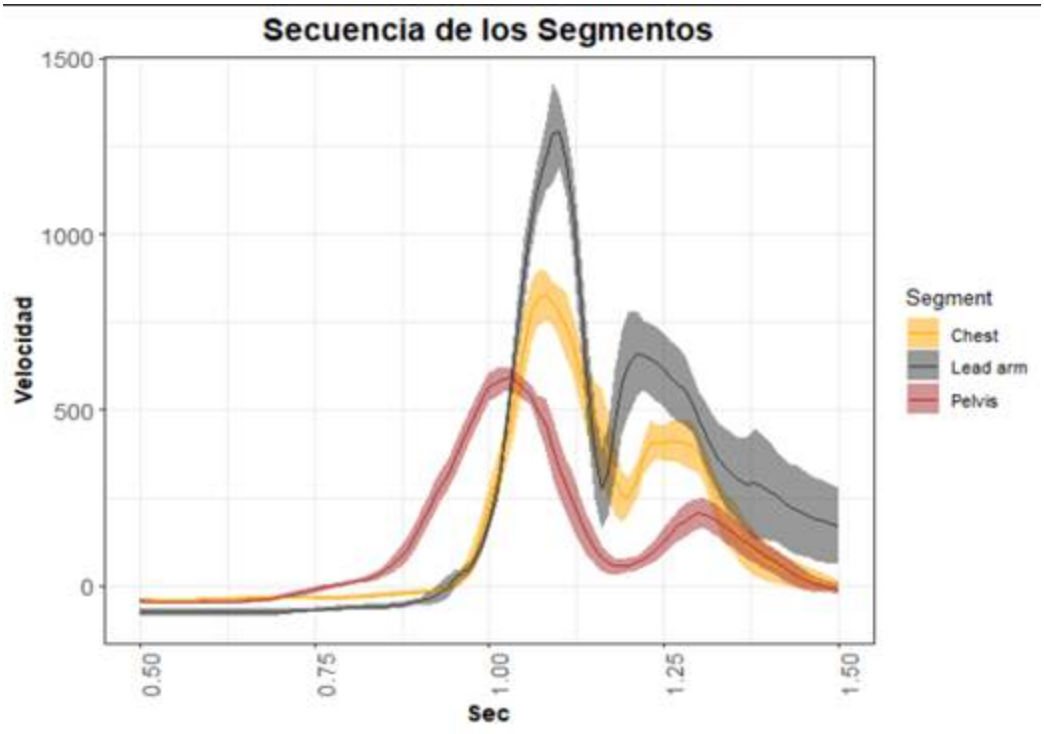
Source: prepared by the author

Lastly, we will look at an example in which we will also visualize the information provided by a device, but using more than one device at a time. These are accelerometers placed on different parts of the body (hip, scapulae and arm) during a sports movement. In this case, we want to assess how the different parts of the body move in relation to the rest during a tennis player's backhand. We could use the summary/calculated results provided by the device, for example, the maximum angular velocity of each of the devices, but we would not be answering the question posed. We want to assess the sequence of the different segments, i.e. which part of the body reaches maximum speed first, which part follows it and which part is the last to reach its maximum speed. Visualizing the raw data signal over the recorded time of the movement of the three accelerometers, we may get the

following graph. This type of kinematic analysis is very common in sports of greater technical complexity.

We see how the peak of the hip is located more to the left than the rest (earlier in time), then the torso gets its peak and finally the arm. Thus, this brings us closer to answering the question. The different peak heights also give us complementary information that can be calculated from the raw data.

Figure 5. Example of information got from more than one device



Source: prepared by the author

These three examples only serve to show the types of analyses we can perform. The most important issue to consider is the objective of our analysis, that is - the question we want to answer. Raw data will be one more tool that we can use to get closer to the solution of our problem.

Considerations in the use of raw data

Data volume

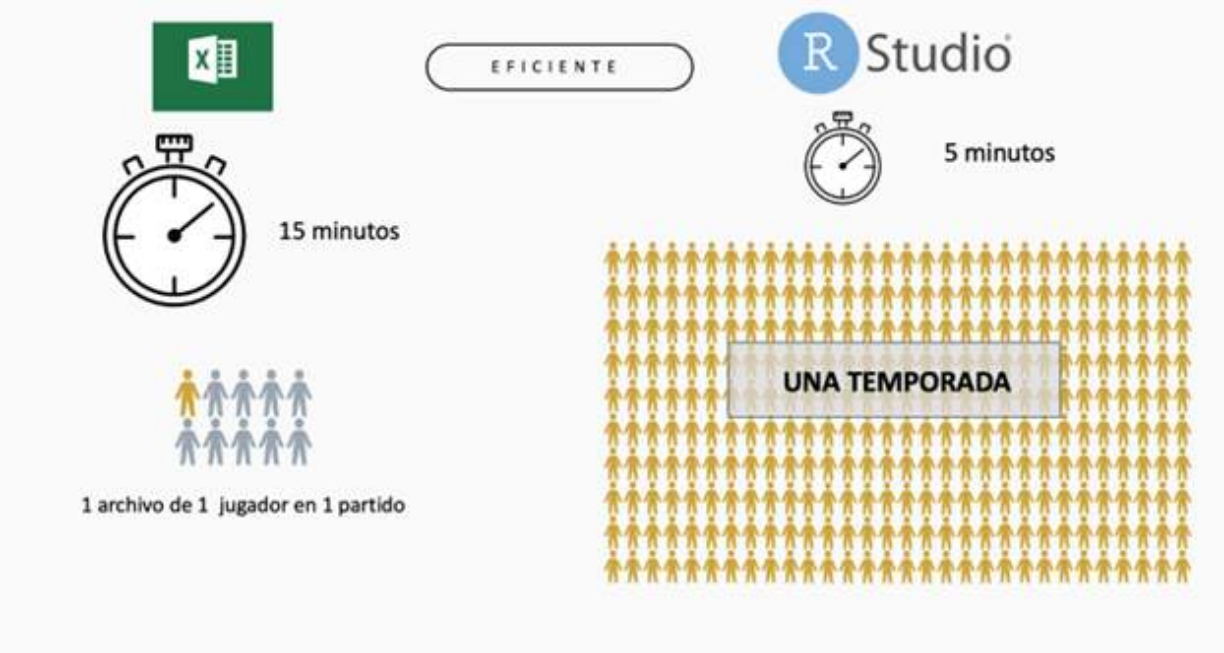
First of all, it is worth noting the high volume of data we work with when we use raw data. We have seen the comparison in jump tests and the difference between the number of observations of an athlete's jump. If we want t carry out this analysis on a daily basis for all the players in our team, the volume of data grows exponentially. In addition, the more variables we want to calculate or updates we want to make to the raw data, the greater the complexity of the data will be.

For this reason, it is necessary to use tools that allow us to be efficient in managing these data structures. RStudio is once again an ideal option to perform the analyses using raw data.

To visualize the capabilities of RStudio compared to more common tools and highlight their differences, we will compare the same analysis process carried out using Excel and RStudio. In this comparison, the aim was to calculate the SubMIPs (example we saw above) for a football team. The same calculations were performed with Excel and RStudio, which followed these steps:

- importing the data into Excel or RStudio;
- filtering the speed and acceleration signal (we will look into this below);
- calculating complementary variables;
- apply moving averages to calculate SubMIPs;
- identifying the SubMIPs;
- calculating the characteristics of each period;
- summarizing the characteristics of the session.

Figure 6. RStudio vs. Excel



Source: prepared by the author

In the diagram we can clearly see how efficient both tools are. While in Excel we need 15 minutes to complete the analysis of a single session for one player due to the need to use multiple Excel workbooks to support the computational demand of calculations, with RStudio we can analyse an entire season in 5 minutes for all the players on the team. The efficiency of RStudio to perform tasks with this type of data is clear.

Self-assessment: The correct option is highlighted in yellow

Which of the following tasks were performed with both Excel and RStudio in the comparative analysis of SubMIPs for a football team?

- [ ] A) Importing the data into Excel or RStudio.
- [ ] B) Filtering the speed and acceleration signal.
- [ ] C) Generating three-dimensional graphics of the players' trajectories.
- [ ] D) Calculating complementary variables.
- [ ] E) Applying moving averages to calculate SubMIPs.

SUBMIT

Filtering data

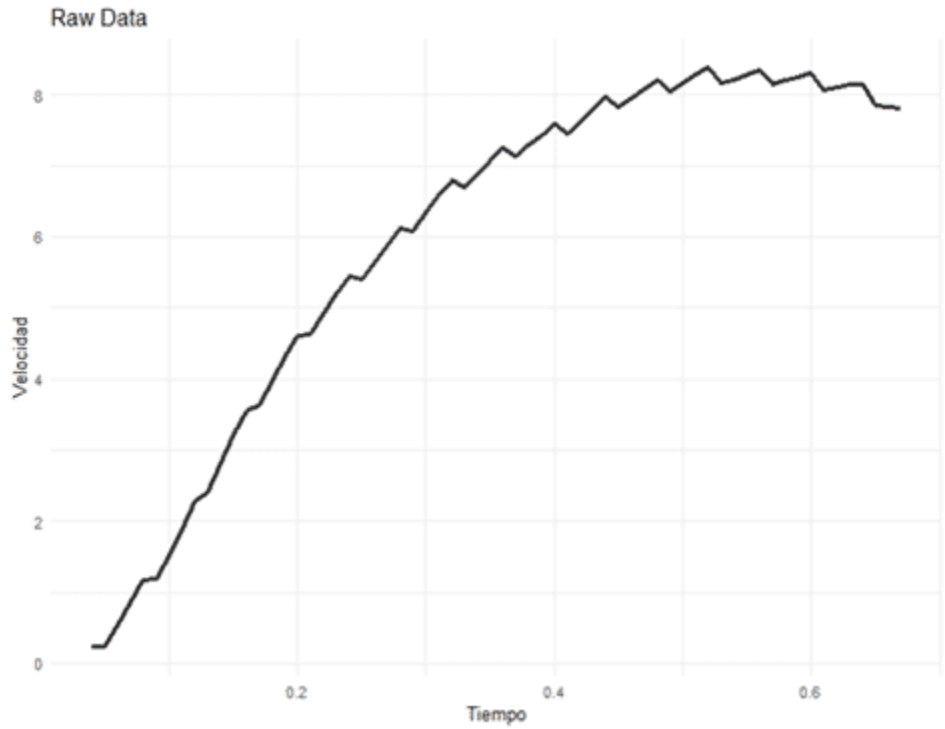
This second aspect is more complex, but of great importance when working with raw data: filtering data. When we work with devices

that record data at high frequency, often the signal we get contains the so-called "noise". This noise refers to small variations in the recorded data that interfere with the quality of the information. This noise can affect the subsequent steps of data processing and analysis, and is common and inherent to the type of signal capture.

To correct that noise and get a valid signal to use, signal filtering methods are used; these allow noise to be reduced without losing the relevant information. It is a complex concept and we can turn to the literature applied to our environment to delve into this content (Hoppe et al., 2018), but in a simplified way we could refer to the application of filtering of a variable such as one that allows noise to be reduced (eliminating unwanted peaks or excessive variations in the data) without losing relevant information (maintaining the shape or trend of the data throughout the time series).

In the following image we can see the speed signal during a sprint recorded by a radar. The signal contains some noise (small increases and decreases over time). If we know the type of movement we are registering (a linear sprint) we can assume that the athlete is moving while increasing the speed (constant positive acceleration). If we see that there are increases and decreases in speed in the signal, it would suggest that the athlete is constantly slowing down and accelerating, which would not be representing the movement we want to register.

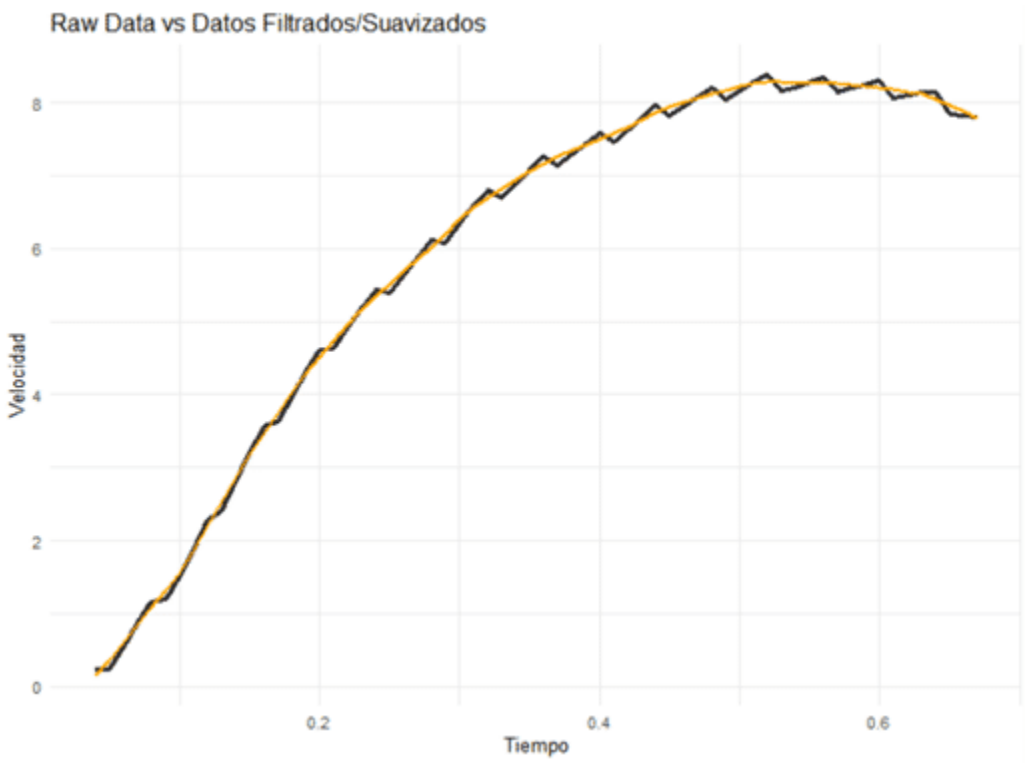
Figure 7. Speed signal during a sprint recorded by a radar



Source: prepared by the author

With the above considerations, we have decided to apply a filter to the speed signal to reduce the noise, while maintaining as much information as possible.

Figure 8. Speed signal during a sprint recorded by a radar – filtering application



Source: prepared by the author

In the image above, we can see the filtered (sometimes also called smoothed) signal (orange line), compared to the raw data (black line). The filtering of the data meets the objective set: it reduces the unnecessary noise of the record by keeping the relevant information, a trend that follows the line.

As we have mentioned before, it is a rather complex procedure since there are multiple types of filters and parameters to regulate the expected result. If we want to replicate different research studies or use the raw data of any of the technologies we generally use (GPS, force platforms, accelerometers), it is essential to look into the

documentation or literature to find out what types of filters they used, since otherwise the results we may get would be different.

To sum up, we can highlight the possibilities and advantages of using raw data in our analyses:

- Detail: it helps in being more precise in assessing the results we get, identifying different phases with greater precision and knowing how the final result has been produced.

- Adaptable/customizable: it allows us to carry out analyses with the preset parameters, offering greater versatility and potential to the possibilities of the analysis.

- Transparency: on many occasions, the methods for processing the data are not described in the literature, so we are limited to knowing where the results come from. Using our own analysis with raw data allows us to have a clearer process and spot possible aspects for improvement.

- Repeatability: the previous point helps us to use the same methods in the future, and to make more accurate comparisons, since it is also common for different companies to make changes in the

processing of data over time, which prevents comparisons with more recent data from being reliable.

CONTINUE

References

Bishop, C., Jordan, M., Torres-Ronda, L., Loturco, I., Harry, J., Virgile, A., Mundy, P., Turner, A., & Comfort, P. (2023). Selecting metrics that matter: comparing the use of the countermovement jump for performance profiling, neuromuscular fatigue monitoring and injury rehabilitation testing. *Strength and Conditioning Journal,* 45(5), 545-553.

Caro, E., Campos-Vázquez, M. Á., Lapuente-Sagarra, M., & Caparrós, T. (2022). Analysis of professional soccer players in competitive match play based on submaximum intensity periods. *PeerJ,* 10, e13309.

García, A., Castellano, J., Mendez-Villanueva, A., Gómez-Díaz, A., Cos, F., & Casamichana, D. (2020). Physical Demands of Ball Possession Games in Relation to the Most Demanding Passages of a Competitive Match. *Journal of Sports Science & Medicine*, 19, 1-9.

Gathercole, R. J., Stellingwerff, T., & Sporer, B. C. (2015). Effect of Acute Fatigue and Training Adaptation on Countermovement Jump Performance in Elite Snowboard Cross Athletes. *Journal of Strength and Conditioning Research*, 29(1), 37-46.

Hoppe, M. W., Baumgart, C., & Freiwald, J. (2018). Estimating external loads and internal demands by positioning systems and innovative data processing approaches during intermittent running activities in team and racquet sports. *Sport-Orthopädie - Sport-Traumatologie - Sports Orthopaedics and Traumatology,* 34, 3-14.