

---

# **ANALISI DI UN DATASET DI SERIE TEMPORALI**

---

**Statistica e Analisi dei dati**

**Autori**

Basilicata Salvatore  
Casaburi Giovanni  
2025

# Contents

<b>1 Il Problema</b>	<b>3</b>
1.0.1 Informazioni del Dataset . . . . .	3
1.0.2 Strutturazione del Dataset Originale . . . . .	3
1.1 Introduzione alle Domande di Ricerca . . . . .	4
<b>2 Data Pre-Processing</b>	<b>5</b>
2.1 Verso la Strutturazione del Dataset . . . . .	5
2.1.1 Rimozione Duplicati e Allineamento Temporale . . . . .	6
<b>3 Analisi</b>	<b>7</b>
3.1 Analisi Univariata . . . . .	7
3.1.1 Distribuzioni per attività e soggetto . . . . .	8
3.1.2 Metriche per tipo di misurazione . . . . .	11
3.1.3 Metriche per attività e soggetto . . . . .	12
3.2 Analisi Bivariata . . . . .	14
3.2.1 Correlazione variabili con Attività . . . . .	14
3.2.2 Cross-Correlation . . . . .	16
3.3 Serie Temporali . . . . .	24
<b>4 Spiegazione Metriche e Metodologie</b>	<b>28</b>
4.0.1 Adjusted Rand Index (ARI) . . . . .	28
4.0.2 V-Measure . . . . .	35
<b>5 Risposta alla Prima domanda di Ricerca</b>	<b>35</b>
<b>6 Risposta alla Seconda Domanda di Ricerca</b>	<b>36</b>
<b>7 Analisi di un Dataset Generato</b>	<b>37</b>
7.1 Generazione del dataset . . . . .	40
7.2 Struttura del Dataset . . . . .	41
7.3 DataPreProcessing . . . . .	41
7.4 Analisi Univariata . . . . .	42
7.5 Analisi Bivariata . . . . .	43
<b>8 Sulla qualità degli LLM nel generare dataset numerici</b>	<b>47</b>
8.1 Task di clustering . . . . .	47
8.1.1 È possibile determinare un'attività da una finestra di dati senza conoscere il paziente? . . . . .	47
8.1.2 È possibile determinare il paziente da una finestra di dati conoscendo l'attività? . . . . .	47
8.2 Statistica inferenziale . . . . .	48

# 1 Il Problema

Il problema che si affronta in questo studio di ricerca è principalmente l'analisi di un dataset contenente le misurazioni di alcuni soggetti mentre svolgono delle attività fisiche. Lo scopo è studiare le feature presenti in questo dataset tramite statistica descrittiva per poi utilizzare le stesse per effettuare attività di clustering o regressione. Una volta condotta l'analisi in questo dataset, si procederà a capire se fornendo l'adeguato prompt ad un Large Language Model, esso riuscirà a generare un dataset sintetico contenente dei dati simili a quelli del dataset originale. In seguito verrà condotto lo stesso studio effettuato sul dataset originale su quello generato dall'LLM, confrontando anche i risultati ottenuti nei vari dataset. Partiremo presentando la struttura del dataset.

## 1.0.1 Informazioni del Dataset

Il dataset selezionato [1] consiste in una collezione di **serie temporali multivariate** registrate da uno smartphone Samsung Galaxy S II posizionato all'altezza della vita di 30 partecipanti (età 19–48 anni).

Per ogni soggetto sono state acquisite misurazioni durante sei attività specifiche:

- Attività relative alla camminata: *Walking*, *Walking Upstairs*, *Walking Downstairs*
- Attività statiche: *Sitting*, *Standing*, *Laying*

I segnali rilevati includono:

- Accelerazione lineare triassiale (assi X, Y, Z)
- Velocità angolare triassiale (assi X, Y, Z)

Altre caratteristiche:

- Frequenza di campionamento: 50 Hz (1 campione ogni 20 ms)
- Totale osservazioni: 508927
- Dispositivo: Accelerometro e giroscopio integrati nello smartphone
- Validazione: Sessioni registrate su video per l'etichettatura precisa delle attività

Le misurazioni combinate producono un insieme di dati multivariato, con sincronizzazione temporale garantita dall'orologio interno del dispositivo.

## 1.0.2 Strutturazione del Dataset Originale

Il dataset presenta una segmentazione in finestre temporali di 2.56 secondi, dove ogni finestra contiene **128 record** (128 campioni × 50Hz). Tra due finestre consecutive appartenenti alla stessa misurazione è presente un **overlap del 50%**. La struttura fornita è organizzata come segue:

- *README.txt*: Descrizione generale del dataset
- *activity\_labels.txt*: Mappatura delle activity labels (es. 1) ai nomi delle attività (es. *WALKING*)

- **Data Files:**

- *Training e Test:*

- \* *X\_train.txt, X\_test.txt*: Vettori delle feature
    - \* *y\_train.txt, y\_test.txt*: Etichette delle attività
    - \* *subject\_train.txt, subject\_test.txt*: ID partecipanti

- **Raw Signal Files** (cartella *Inertial Signals*):

- \* Accelerometro triassiale:

- *body\_acc\_x\_train.txt* (asse X)
      - *body\_acc\_y\_train.txt* (asse Y)
      - *body\_acc\_z\_train.txt* (asse Z)

- \* Giroscopio triassiale:

- *body\_gyro\_x\_train.txt* (velocità angolare asse X)
      - *body\_gyro\_y\_train.txt* (velocità angolare asse Y)
      - *body\_gyro\_z\_train.txt* (velocità angolare asse Z)

## 1.1 Introduzione alle Domande di Ricerca

Le domande di ricerca sono le seguenti:

- **Attività fisica:** È possibile riconoscere in modo affidabile l'attività fisica (ad es. camminata, corsa, seduta) analizzando una finestra arbitraria che comprende velocità angolari e accelerazioni lineari? In altre parole, i dati raccolti in un intervallo così breve contengono pattern discriminanti sufficienti per distinguere le diverse attività?
- **Identificazione del partecipante:** Riusciamo a identificare in maniera accurata un partecipante basandoci su una finestra di misurazioni consecutive del suo segnale (contenente velocità angolari e accelerazioni lineari)? E come varia la capacità di identificazione se, oltre ai dati sensoriali, integriamo l'informazione relativa all'attività fisica svolta in quella finestra?
- I Large Language Model sono strumenti adatti per la generazione di dati sintetici contenenti accelerazioni lineari e velocità angolari?
- Le feature generate dai Large Language Model possono essere ricondotti ad una distribuzione normale?
- I dati generati dai Large Language Model possono sostituire i dati reali per le attività di clustering?
- Come si differiscono le metriche dei dati sintetici generati dagli Large Language Model dalle metriche dei dati reali?

## 2 Data Pre-Processing

In questa sezione descriveremo i passaggi che, a partire da un insieme di dati non strutturati in formato `.txt`, ci hanno permesso di creare un dataset strutturato e codificato in formato `.csv`, pronto per le fasi successive dell'analisi.

### 2.1 Verso la Strutturazione del Dataset

I file di testo separati (`body_acc_x.txt`, `body_acc_y.txt`, ecc.) sono stati unificati e concatenati insieme ai file `subject.txt` e `y.txt` (che indica l'attività corrispondente). Il dataset risultante, in formato `.csv`, presenta le seguenti colonne:

- `Subject`
- `Activity`
- `body_acc_x_1, ..., body_acc_x_128`
- `body_acc_y_1, ..., body_acc_y_128`
- `body_acc_z_1, ..., body_acc_z_128`
- `body_gyro_x_1, ..., body_gyro_x_128`
- `body_gyro_y_1, ..., body_gyro_y_128`
- `body_gyro_z_1, ..., body_gyro_z_128`
- `total_acc_x, total_acc_y, total_acc_z`

In totale, il dataset presenta 1154 colonne (2 colonne principali (`Subject,Activity`) più 9 gruppi di 128 colonne).

Successivamente, per ogni riga originale, sono state generate nuove righe basate sulle 128 finestre. Per ciascuna finestra sono stati aggiunti gli attributi `WindowID` e `Offset`, creando una nuova struttura con le seguenti colonne:

- `Subject`
- `Activity`
- `WindowID`
- `Offset`
- `body_acc_x, body_acc_y, body_acc_z`
- `body_gyro_x, body_gyro_y, body_gyro_z`
- `total_acc_x, total_acc_y, total_acc_z`

### 2.1.1 Rimozione Duplicati e Allineamento Temporale

Dopo aver strutturato il dataset, abbiamo preso in considerazione due aspetti fondamentali:

- Ogni record appartiene a una finestra temporale di 128 record.
- Due finestre consecutive presentano un overlap del 50%.

L'overlap del 50% potrebbe influenzare l'analisi. In ogni misurazione continua, tutti i valori risultano duplicati, ad eccezione dei primi 64 e degli ultimi 64 record di ciascuna finestra. Questo aspetto può introdurre ridondanza nei dati e, potenzialmente, influire sull'accuratezza dei risultati. Per garantire un'analisi più approfondita e una rappresentazione fedele delle serie temporali, è cruciale avere una concezione chiara delle finestre effettive. Abbiamo quindi deciso di unificare le finestre consecutive che presentano overlap in un'unica finestra consecutiva, rimuovendo gli overlap (Tabella 1).

WindowID Prima	1 <sup>a</sup> metà duplicata?	Offset Originale	Nuovo Offset	Nuovo WindowID
79	No	1-128	1-128	79
80	Si	1-128	129-192	79
81	Si	1-128	193-256	79
82	No	1-128	1-128	80

Table 1: Esempio della tecnica di rimozione di overlap utilizzata

Questo approccio garantisce che ci sia continuità temporale senza gap o sovrapposizioni e che si conservi la relazione tra Offset e timestamp reale. Inoltre, ciò garantisce anche una riduzione del 50% nel numero di finestre pur mantenendo tutta l'informazione originale.

### 3 Analisi

Per analizzare il dataset procederemo prima con un'analisi univariata classica, e successivamente con un'analisi propria per le serie temporali, ed infine con un'analisi bivariata.

#### 3.1 Analisi Univariata

Prima di rispondere alle domande di ricerca, è fondamentale capire se i dati contengano sufficienti informazioni discriminative. L'analisi univariata ci permette di scovare due tipi di problemi: 1) dati troppo rumorosi che rendano impossibile qualsiasi separazione 2) pattern distributivi che richiedano preprocessing specifico.

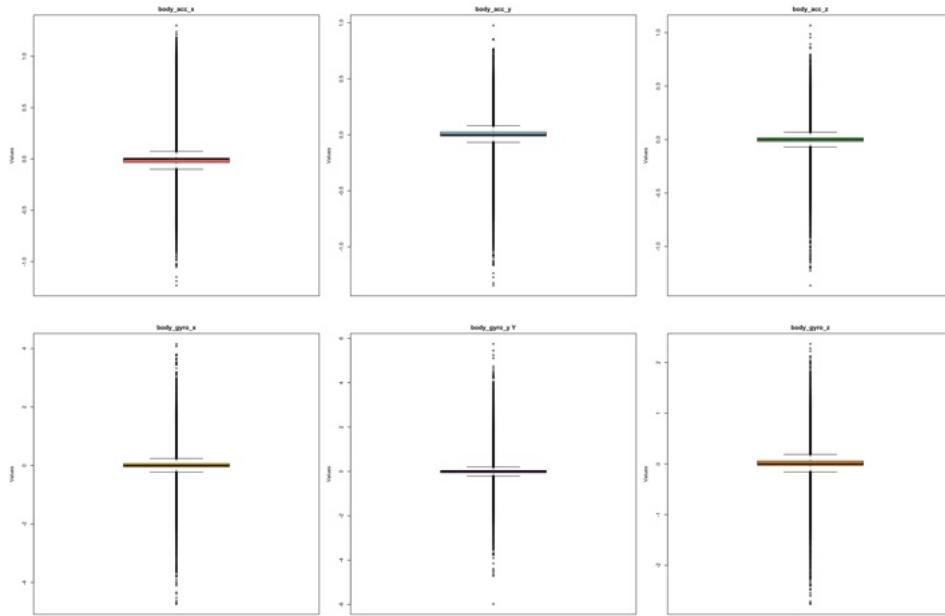


Figure 1: Boxplot dei sensori: i pallini rossi mostrano oltre 150.000 outlier per variabile secondo la regola  $1.5 \times \text{IQR}$ . Un paradosso apparente, considerando che media e mediana sono quasi nulle.

Questo esercizio di "data archeology" inizia dai boxplot (Fig. 1). A colpo d'occhio, notiamo un'esplosione di outlier che sembrerebbero invalidare qualsiasi analisi. Ma è davvero rumore? O stiamo applicando criteri sbagliati? Per districarci, passiamo al setaccio le metriche nella Tabella 2.

Table 2: Metriche riassuntive - La diagnosi delle code

Sensore	Media	Mediana	SD	Skewness	Kurtosis	Min	Max	MAD
body_acc_x	-0.001	-0.001	0.198	1.009	4.506	-1.232	1.300	0.024
body_acc_y	0.000	0.001	0.124	-0.990	5.938	-1.345	0.976	0.025
body_acc_z	0.000	0.000	0.109	-0.498	6.932	-1.365	1.067	0.026
body_gyro_x	0.001	0.000	0.416	-0.223	5.882	-4.734	4.155	0.086
body_gyro_y	0.000	-0.001	0.386	0.553	9.640	-5.974	5.746	0.077
body_gyro_z	0.000	0.001	0.260	-0.519	5.812	-2.763	2.366	0.063

I numeri raccontano una storia diversa: le altissime kurtosis (fino a 9.64) tradiscono distribuzioni a picco stretto con code grasse. Non sono outlier veri, ma estremi fisiologici! Un accelerometro a riposo avrà misurazioni vicine allo zero (MAD 0.025), ma durante un movimento brusco picchi fino a  $\pm 5g$ (*body\_gyro\_y*).

Questo spiega perché il 30% dei dati superi  $1.5 \times \text{IQR}$ : la regola boxplot assume normalità, ma i nostri dati hanno code iper-trofiche. Per non tagliare informazioni preziose, abbiamo considerato come outlier solo i valori oltre  $3\sigma$  (Fig. 2, 15).

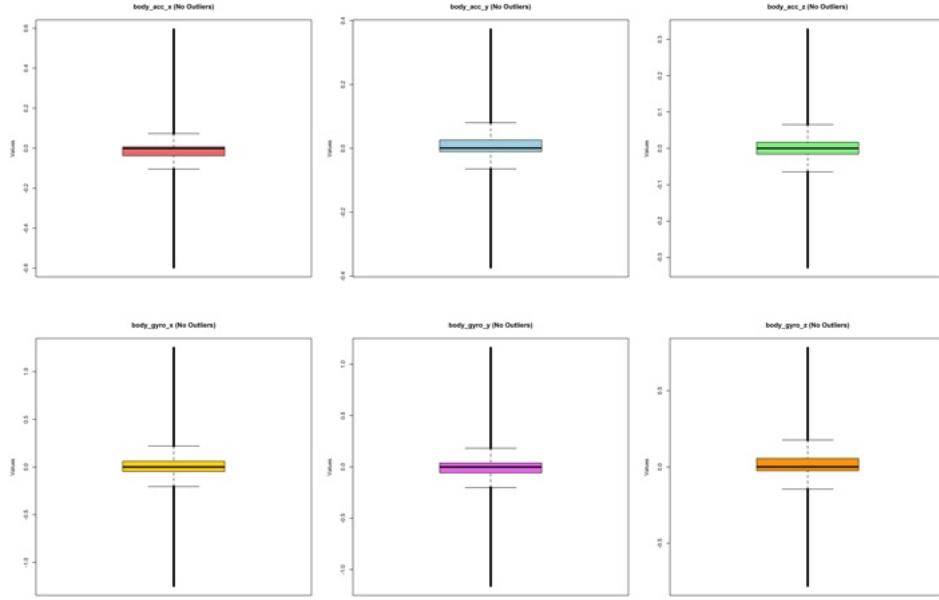


Figure 2: Boxplots delle variabili senza outliers.

Come anche visibile dal grafico delle distribuzioni (Fig4), la coesistenza di picchi stretti (MAD 0.025) e ampi range dinamici (max-min fino a 11.72 in *body gyro y*) implica che i segnali alternano fasi quiescenti (dominanti in frequenza) a brevi transienti ad alta energia, tipici dei pattern motori naturali.

### 3.1.1 Distribuzioni per attività e soggetto

Dopo aver visualizzato le distribuzioni in generale per ogni feature, analizziamo come esse si comportano nella singola attività tra i vari soggetti:

Guardando i Kernel density plot (Fig. 5) e i boxplot (Fig. 6) per ogni feature nell'attività di camminata, notiamo innanzitutto che le distribuzioni sono più interpretabili rispetto all'analisi generale. Ogni feature nell'attività di camminata vediamo che ogni soggetto si differisce nell'attività ad esclusione del giroscopio delle y dove i soggetti hanno delle distribuzioni simili. Dai boxplot delle feature notiamo che il valore mediano cambia in ogni soggetto nelle accelerazioni, invece nei giroscopi il valore mediano è molto simile tra i soggetti. Il grafico con la linea nera indica come si distribuisce in generale la feature nell'attività e notiamo che alcune feature si approssimano ad una normale mentre altre presentano delle code più lunghe a destra o a sinistra.

La seconda serie di grafici , (Fig. 7, 8), mostra come si comportano le feature nell'attività

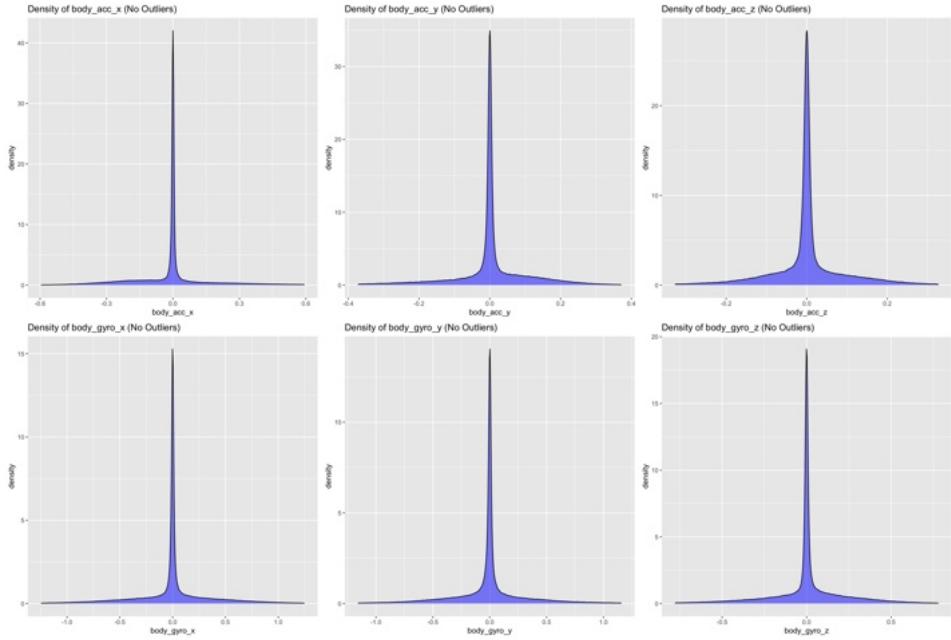


Figure 3: Distribuzione delle variabili senza outliers.

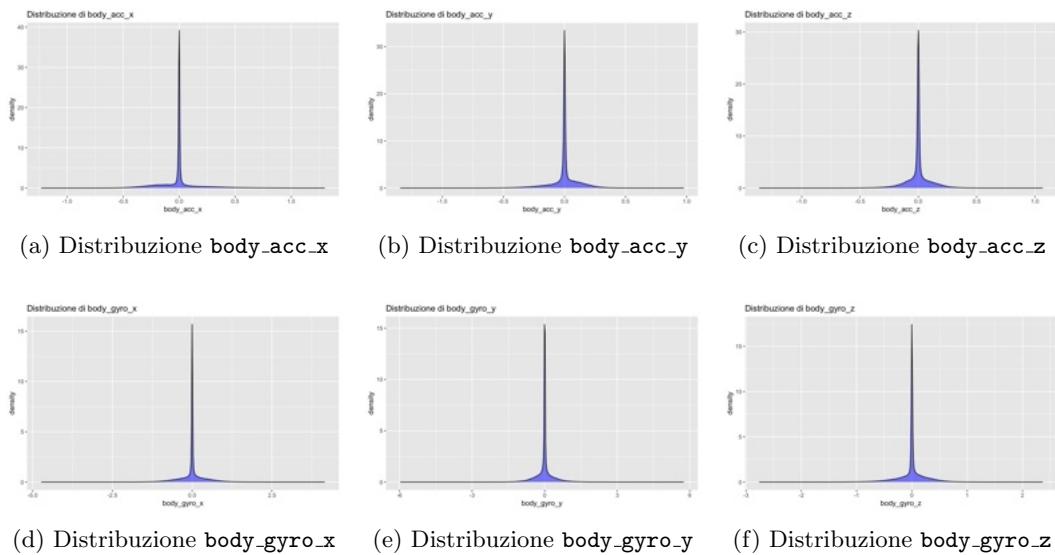


Figure 4: Distribuzioni marginali: picchi strettissimi (SD 0.1-0.4) ma code che arrivano a  $\pm 5$  unità. Sono proprio queste code, legate ai transienti motori, a contenere le firme cinematiche per distinguere attività e pazienti.

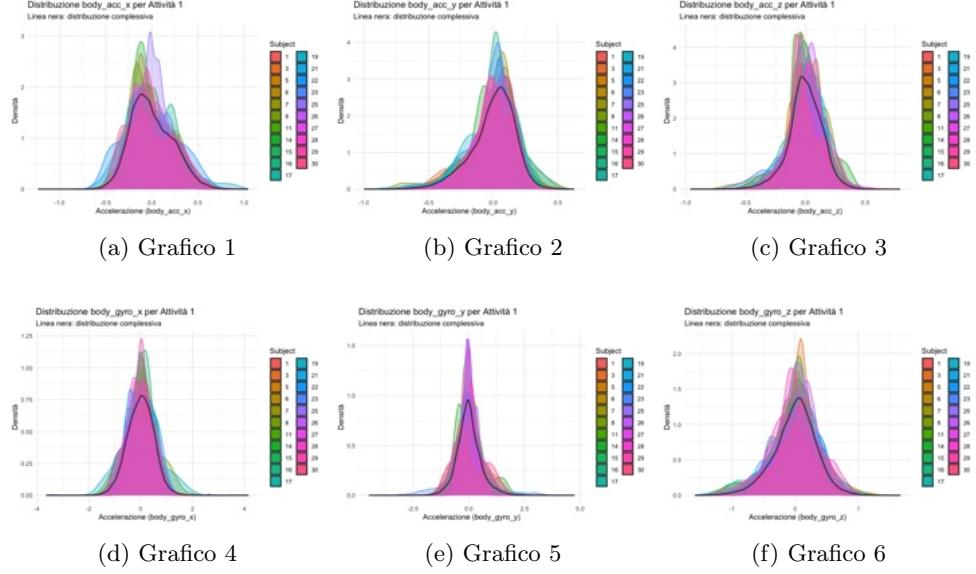


Figure 5: Distribuzione delle feature per camminata tra i soggetti

di salire le scale e notiamo che l’accelerazione sull’asse x per ogni soggetto si distribuisce molto diversamente tra i vari soggetti, quindi nel momento di classificazione sarà una feature chiave che distinguerà i soggetti. Le altre feature si comportano in modo simile tra i vari soggetti, ad eccezione del giroscopio sull’asse Z, in cui si osserva un soggetto che si distingue nettamente dagli altri, probabilmente poiché si comporta come un outlier. Guardando i boxplot, notiamo che l’IQR è simile per le accelerazioni e per i giroscopi; inoltre, per il giroscopio sull’asse Y il valore mediano è uguale nella maggior parte dei soggetti.

Per l’attività di scendere le scale (Fig. 9 e Fig. 10) si osserva nuovamente che l’accelerazione lungo l’asse X è quella che differenzia maggiormente i soggetti che eseguono l’attività, mentre le altre accelerazioni tendono a sovrapporsi notevolmente. Lo stesso ragionamento si applica ai giroscopi, per i quali le distribuzioni tra i vari soggetti si sovrappongono considerevolmente. Dai boxplot emerge, inoltre, che la mediana dell’accelerazione lungo l’asse X per ogni soggetto, rispetto alle attività precedenti, è sempre negativa, mentre in precedenza tendeva a essere pari a zero. L’IQR per l’accelerazione lungo gli assi Y e Z è diminuito drasticamente rispetto alle attività precedenti, indicando che queste feature presentano una varianza minore (escludendo gli outlier) rispetto a quanto osservato in precedenza. I giroscopi, invece, continuano a mostrare una distribuzione simile a quella riscontrata nei boxplot precedenti.

L’ultima serie di grafici (Fig 11 e Fig 12) mostra la distribuzione delle feature nella prima attività stazionaria: stare seduti. Notiamo che, per tutti i soggetti, le varie feature presentano distribuzioni simili, salvo alcune piccole differenze non ben evidenti dai grafici, dovute a micromovimenti eseguiti dai soggetti. Dai boxplot si evince che, in queste feature e in questa attività, molte misurazioni sono considerate outlier, poiché nell’attività di stare fermi ci si aspetterebbe di ottenere sempre lo stesso valore, dato che il soggetto non esegue alcun movimento. Tuttavia, tali misurazioni, pur venendo mostrate come outlier dai boxplot, in realtà non lo sono, soprattutto se si considerano le feature per ciascun soggetto, poiché sono proprio questi micromovimenti o movimenti bruschi a permetterci di distinguere quale

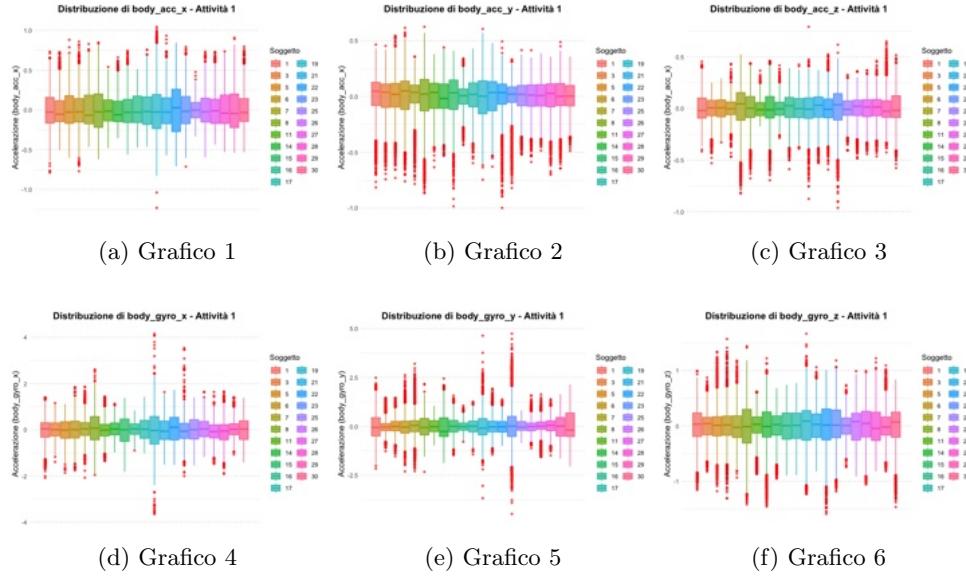


Figure 6: Distribuzione delle feature per camminata tra i soggetti

soggetto stia eseguendo l'attività in esame.

### 3.1.2 Metriche per tipo di misurazione

Proseguendo l'analisi univariata, esaminiamo ora le metriche riassuntive per attività e asse riportate nella Tabella 3. Questi indicatori, che comprendono kurtosi, media, mediana, deviazione standard, skewness e varianza, offrono ulteriori spunti per comprendere la struttura intrinseca dei dati registrati in condizioni diverse.

#### Osservazioni principali:

- Concentrazione e dispersione:** Le attività 1, 2 e 3 presentano valori di kurtosi relativamente moderati (da circa 2.8 a 6.15) e deviazioni standard consistenti (da 0.14 fino a 0.38). Questo suggerisce distribuzioni con una concentrazione attorno al valore centrale, ma con una dispersione sufficiente a catturare la variabilità del movimento. Al contrario, le attività 4, 5 e 6 mostrano kurtosi estremamente elevate (fino a 110 per l'asse  $x$  nell'attività 6), segnalando distribuzioni fortemente leptocurtiche, ovvero segnali per lo più stabili, con variazioni minime (SD dell'ordine di centinaia o millesimi) interrotte da transitori sporadici.
- Centrale simmetria:** Medie e mediane risultano sostanzialmente nulle in ogni condizione, confermando una centralità dei dati, indipendentemente dall'attività o dall'asse considerato. Ciò indica che, sebbene i segnali subiscano forti transitori, il loro centro rimane invariato.
- Asimmetria (skewness):** La skewness evidenzia asimmetrie interessanti: per esempio, mentre nell'attività 1 l'asse  $y$  presenta una skewness negativa (circa -0.76), l'asse  $x$  risulta leggermente asimmetrico in senso opposto (0.40). In casi estremi, come l'attività 6, l'asse  $x$  assume uno skewness fortemente negativo (-7.81), mentre l'asse

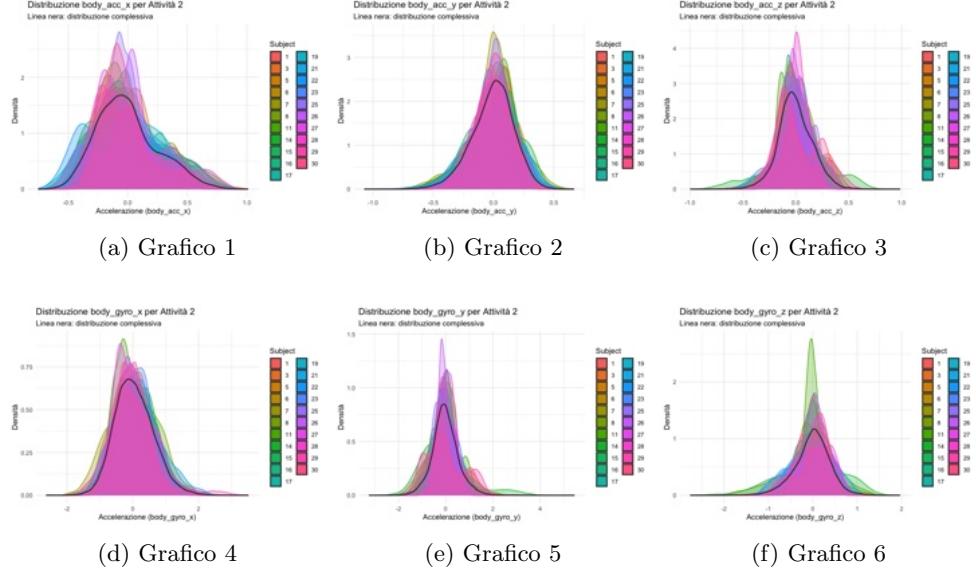


Figure 7: Distribuzione delle feature per camminata a salire tra i soggetti

$z$  mostra una marcata asimmetria positiva (3.09). Tali differenze indicano che, durante alcune attività, eventi rari e intensi (probabilmente legati a bruschi movimenti o anomalie momentanee) possono influenzare fortemente la forma della distribuzione.

- **Varianza e robustezza del segnale:** L'ampia differenza nelle deviazioni standard e varianze tra le attività dinamiche (1-3) e quelle caratterizzate da segnali più stabili (4-6) suggerisce che, mentre in alcune condizioni il movimento introduce una variabilità significativa, in altre il segnale rimane confinato in un range molto ristretto, evidenziando la presenza di transitori isolati.

Questi risultati rafforzano il concetto che la definizione tradizionale di outlier (basata sulla regola  $1.5 \times \text{IQR}$ ) potrebbe risultare troppo restrittiva in presenza di distribuzioni con code pesanti. Infatti, le estremità evidenziate in alcune attività non sono necessariamente rumore, ma possono rappresentare eventi fisiologicamente rilevanti. Per questo motivo, in analogia a quanto osservato nei boxplot, si è ritenuto opportuno adottare soglie più flessibili (ad esempio, basate su 3) per preservare le informazioni discriminanti e consentire un preprocessing mirato.

### 3.1.3 Metriche per attività e soggetto

In questa parte mostreremo per ogni attività, come variano le metriche tra i tipi di misurazione. Per un'analisi più dettagliata, abbiamo raggruppato per attività, per poi calcolare le metriche per ogni soggetto.

Guardando la prima serie di grafici, (Fig. 20), notiamo delle particolarità nelle misure di alcuni sensori, in particolare quelli che riguardano l'accelerazione totale. La media di questi ultimi per l'asse delle x è poco dispersa tra i vari soggetti, stabilendosi verso il valore mediano. Guardando il grafico della deviazione standard dell'accelerazione totale, notiamo

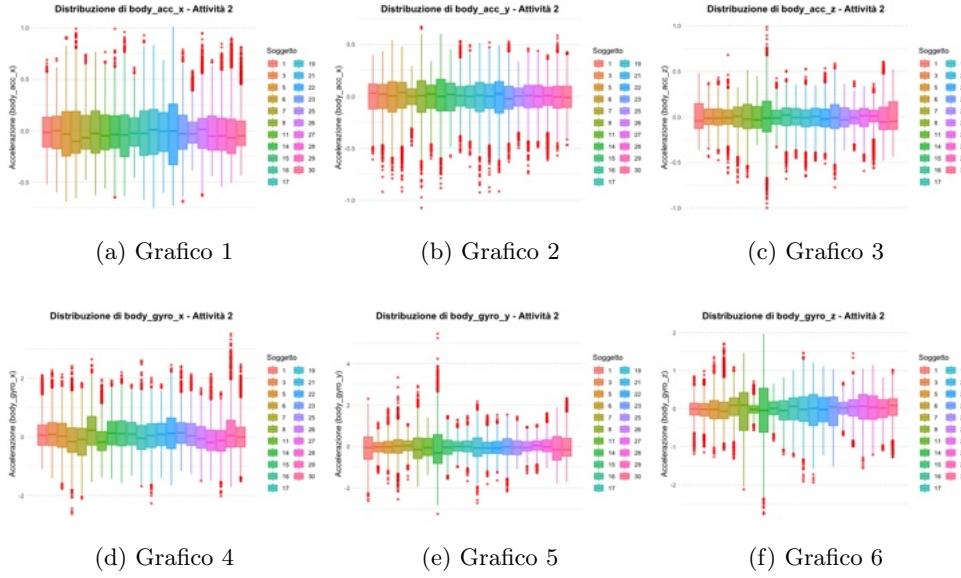


Figure 8: Boxplot per camminata a salire per soggetto

Table 3: Metriche riassuntive per attività e assi

Activity	Asse	Kurtosis	Mean	Median	SD	Skewness
1	x	2.93	-0.00041	-0.02901	0.22747	0.39576
1	y	4.06	-0.00089	0.02325	0.17109	-0.75655
1	z	4.35	-0.00047	0.00173	0.14358	-0.43401
2	x	2.95	-0.00423	-0.03443	0.25909	0.50087
2	y	3.52	-0.00716	0.00681	0.17629	-0.42483
2	z	4.07	-0.00479	-0.01409	0.16555	0.04734
3	x	2.82	0.00245	-0.10378	0.37876	0.75833
3	y	5.01	0.00128	0.02897	0.18957	-0.91712
3	z	6.15	0.00197	0.01470	0.16019	-0.82636
4	x	94.19	-0.00097	-0.00009	0.01756	-0.09724
4	y	63.02	0.00380	0.00051	0.03748	5.31637
4	z	57.58	0.00069	0.00012	0.03512	0.37789
5	x	66.97	0.00069	0.00006	0.01230	2.15627
5	y	43.83	0.00126	0.00005	0.02391	2.85132
5	z	62.58	0.00107	-0.00012	0.02791	-0.13666
6	x	110.87	-0.00272	-0.00032	0.04425	-7.80617
6	y	80.57	-0.00036	-0.00018	0.04913	-0.41718
6	z	65.26	0.00081	0.00019	0.04406	3.08594

che nell'attività di camminata i valori dell'accelerazione totale si disperdonano come i valori dell'accelerazione normale ad eccezione dell'asse z che si differenzia per il valore mediano delle deviazioni standard.

I grafici dell'attività di salita, (Fig. 21) mostrano che la media dell'accelerazione di questa attività è molto simile alla media dell'accelerazione totale della camminata, inoltre, la deviazione standard dell'accelerazione totale continua ad essere uguale in media all'accelerazione normale senza gravità.

Nell'attività di scendere le scale, (Fig. 22), invece troviamo ulteriori conferme sulle devi-

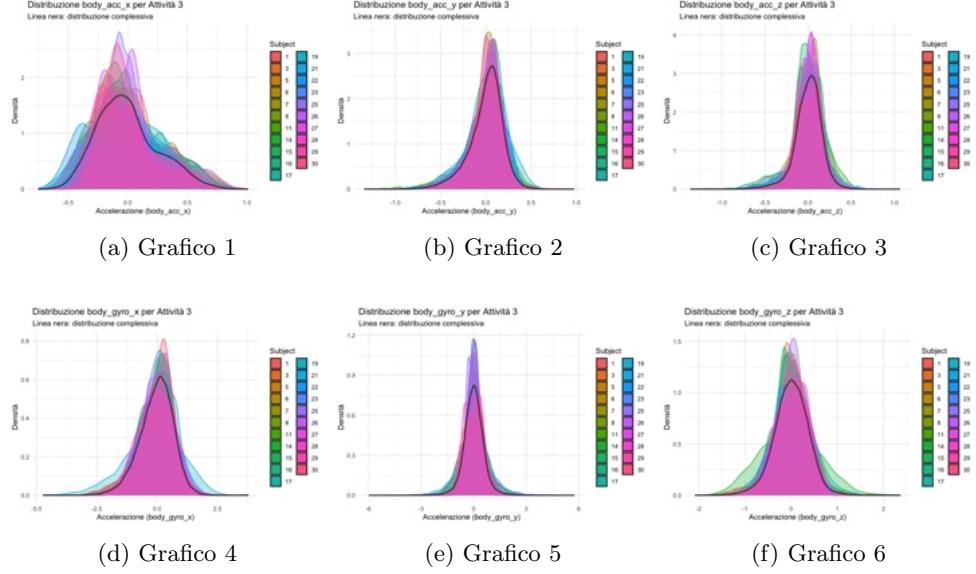


Figure 9: Distribuzione delle feature per camminata a scendere tra i soggetti

azioni standard delle accelerazioni totali tra i soggetti, come nei casi precedenti sono molto simili alle deviazioni standard dell'accelerazione e a differenza delle due precedenti attività, le medie delle accelerazioni totali rispetto all'asse delle x si stabilizzano sul loro valore mediano, presentando solo due outlier. Considerando quindi solo le attività di camminata, notiamo che in generale le metriche dei sensori dei giroscopi e delle accelerazioni in ogni soggetto tendono a cambiare range di variabilità.

Per quanto riguarda le ultime attività, (Fig. 23, (Fig. 24) e (Fig. 25), che sono quelle che non richiedono movimento, notiamo che l'accelerazione dell'asse x e delle z presentano medie che sono quasi sempre nulle, ad eccezione di alcuni soggetti che hanno una leggera accelerazione. L'accelerazione totale delle x nelle attività 4 e 5 ha medie positive prossime all'uno, nonostante non venga rilevata un'accelerazione con il sensore delle x.

### 3.2 Analisi Bivariata

Lo studio dell'analisi bivariata si è concentrato sul capire come ogni variabile è correlata con l'**attività**. Per ogni coppia (variabile, target) è stato creato un grafico per esplorare il tipo di relazione che sussiste tra la variabile e il target ed è stato calcolato il coefficiente di correlazione. Quest'ultimo è stato calcolato utilizzando il metodo di Spearman che non fa assunzioni sul tipo di relazione tra le due variabili considerate, ma solo che la loro relazione può essere descritta con una funzione monotona. Di seguito è illustrata l'analisi bivariata, considerando come target l'attività.

#### 3.2.1 Correlazione variabili con Attività

Come già anticipato nella sezione precedente, per ogni coppia è stato creato uno scatterplot, (Fig. 26, 27, 28), per visualizzare la relazione tra la coppia ed è stato calcolato il coefficiente di correlazione. Dai grafici si nota che la relazione che sussiste tra le variabili con l'attività non è chiara a primo impatto perché i punti tendono a sovrapporsi molto dato che il valore

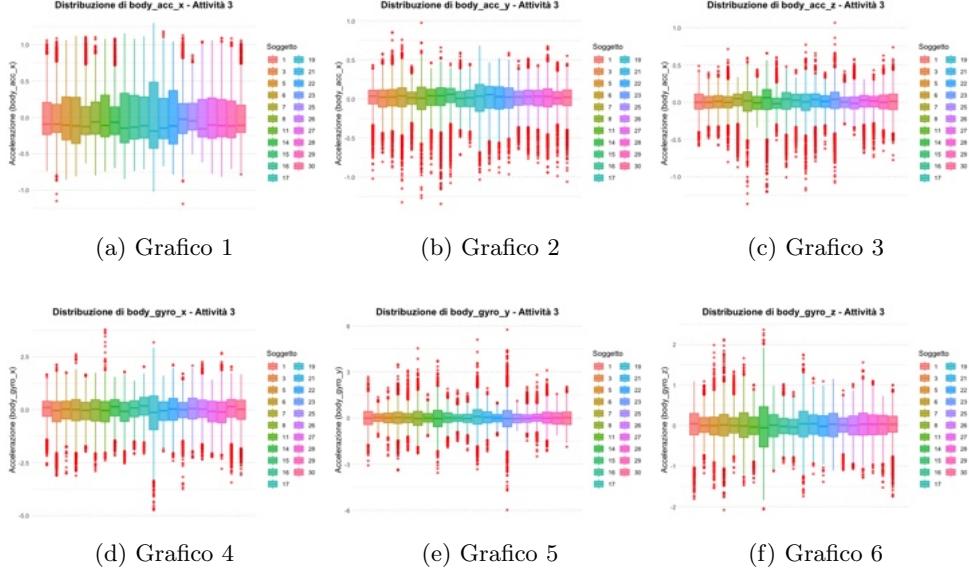


Figure 10: Boxplot feature per camminata a scendere tra i soggetti

tra una misura e l'altra varia di poco, formando delle linee orizzontali allo stesso livello del valore dell'attività, non permettendo di capire a fondo il tipo di relazione che sussiste. Una delucidazione in più e molto più chiara ci è data dal coefficiente di correlazione. La tabella 7 mostra la correlazione di ogni variabile con l'attività. Il coefficiente di Spearman riporta un valore compreso tra  $[-1,1]$ ; se è positivo vuol dire che al crescere di X anche la variabile Y cresce, se è negativo al crescere di X, la variabile Y decresce. Guardando i vari coefficienti notiamo che l'accelerazione dell'asse z ha un coefficiente pari a 1, il che vuol dire che al crescere o al diminuire dell'accelerazione, l'attività fa lo stesso. Se consideriamo che le prime tre attività sono le camminate e le ultime tre dei modi di stare fermi, capiamo che la **camminata normale** ha un'accelerazione sull'asse delle z minore rispetto a quella dell'attività di **salire** e **scendere**. Le altre variabili che presentano una correlazione da tenere in conto sono l'accelerazione totale dell'asse delle x, y e z. L'accelerazione totale delle x è correlata negativamente con l'attività, quindi quando l'accelerazione aumenta vuol dire che l'attività diminuisce, quindi se l'accelerazione sta aumentando vuol dire che il soggetto sta iniziando a camminare, se diminuisce è perchè si sta fermando.

Variabile	Spearman
Body_Acc_X	0.087
Body_Acc_Y	-0.073
Body_Acc_Z	1.00
Body_Gyro_X	0.009
Body_Gyro_Y	0.027
Body_Gyro_Z	-0.011
Total_Acc_X	-0.41
Total_Acc_Y	0.54
Total_Acc_Z	0.26

Table 4: Coefficiente di correlazione tra le variabili e l'Attività

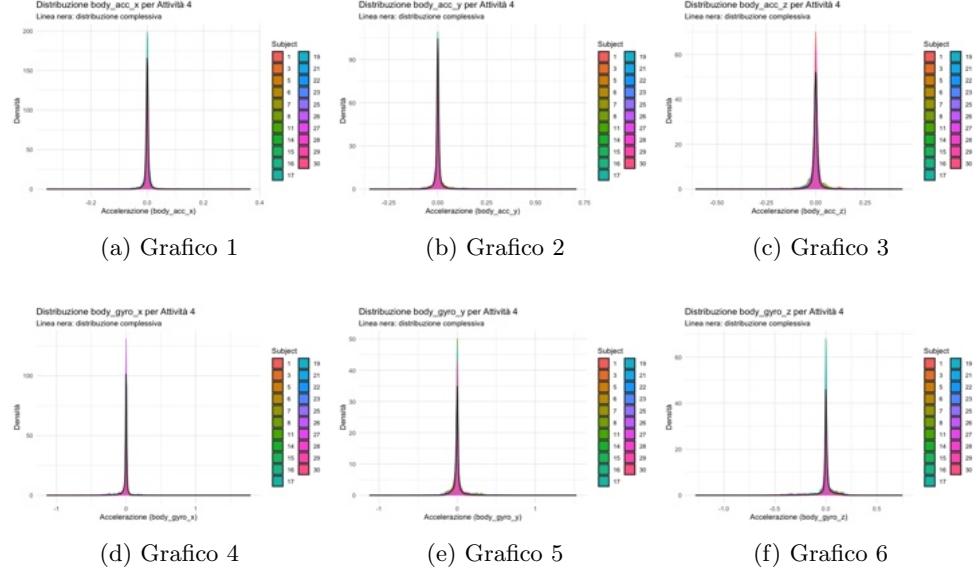


Figure 11: Distribuzione delle feature per stare seduti tra i soggetti

### 3.2.2 Cross-Correlation

In questa parte si è effettuata un’analisi per capire le variabili come sono correlate tra loro nel tempo. Nell’effettuare quest’analisi si sono prese le coppie più rilevante delle feature disponibili:

- Accelerazione - Giroscopio;
- Accelerazione - Totale Accelerazione;
- Accelerazione X - Accelerazione Y;
- Accelerazione Y - Accelerazione Z;
- Accelerazione X - Accelerazione Z;
- Totale Accelerazione X - Totale Accelerazione Y;
- Totale Accelerazione Y - Totale Accelerazione Z;
- Totale Accelerazione X - Totale Accelerazione Z

Per ogni coppia considerata è stato creato un grafico che mostra il Cross-Correlation Coefficient per ogni finestra presente nel dataset. In questi grafici, sull’asse delle x sono indicati i lag temporali, mentre sull’asse delle y sono riportati i valori della correlazione. Sono inoltre tracciate due linee blu tratteggiate che indicano i limiti di significatività statistica: un picco che supera tale limite viene considerato significativo per quel lag.

La prima serie di grafici mostra la cross correlation tra l’accelerazione sull’asse x e quella sull’asse z. La prima osservazione è la presenza di valori speculari, che risultano poi assenti nelle ultime due finestre. Ciò indica che le variabili hanno oscillazioni che seguono una relazione sinusoidale, poiché il corpo accelera e decelera in direzioni opposte in modo regolare. Inoltre, la presenza di picchi a lag positivi suggerisce che, in alcuni momenti del movimento,

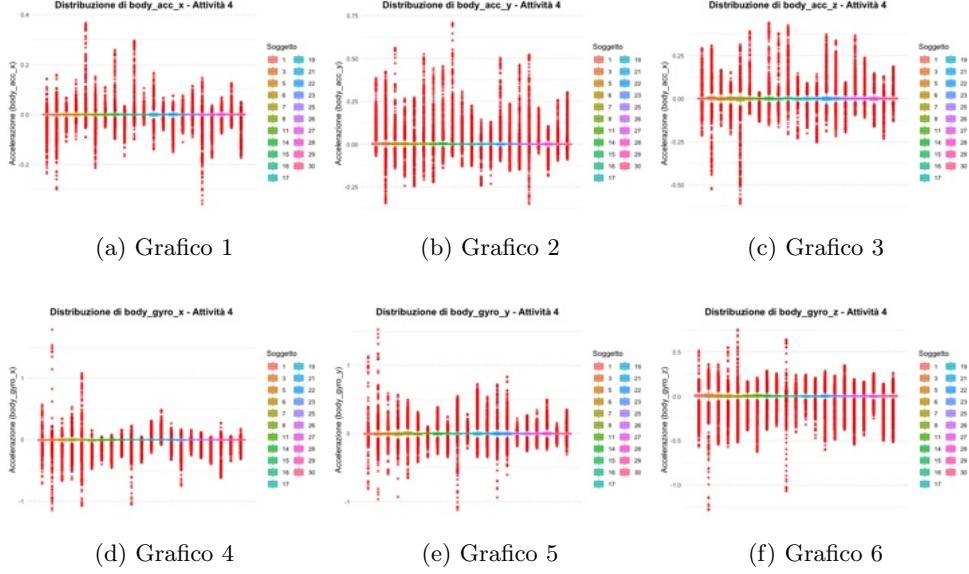


Figure 12: Boxplot feature per attività seduti tra i soggetti

la prima variabile precede la seconda, mentre nella parte finale è la seconda a precedere la prima. Dato che le misurazioni variano notevolmente in alcune finestre, possiamo dedurre che la relazione tra le due variabili non è stabile, ma dipende dall'attività svolta.

Le successive due serie di grafici illustrano la correlazione tra l'accelerazione lungo l'asse x e quella lungo l'asse y, e infine tra l'asse y e l'asse z. Esaminando tali grafici si evince che valgono le stesse considerazioni fatte per la correlazione tra l'asse x e l'asse z.

Successivamente, sono state analizzate le coppie che riguardano le accelerazioni e i rispettivi giroscopi, a partire dalla correlazione tra l'accelerazione sull'asse x e il giroscopio sull'asse x. Osservando questa serie di grafici, si nota innanzitutto che le misurazioni variano da una finestra all'altra, il che suggerisce che la relazione non è stabile, ma è influenzata dal tipo di attività. Un'altra caratteristica osservata è che, nella prima finestra, la seconda variabile precede la prima, come evidenziato dal numero maggiore di picchi a lag negativi rispetto a quelli a lag positivi; successivamente, la relazione tende a stabilizzarsi.

La seconda serie di grafici riguarda la correlazione tra l'accelerazione lungo l'asse y e il giroscopio sull'asse y. In questo caso, nelle prime finestre i grafici risultano molto simili tra loro e presentano sia picchi a lag negativi sia a lag positivi, mentre nelle finestre successive i valori divergono. Ciò indica che la relazione dipende effettivamente dall'attività svolta, nonostante una parte iniziale delle misurazioni sia identica.

L'ultima serie di grafici dei giroscopi esamina la correlazione tra l'accelerazione lungo l'asse z e il giroscopio sull'asse z, evidenziando che anche in questo caso la relazione varia in funzione dell'attività.

La penultima serie di grafici analizza la correlazione tra le accelerazioni lungo un asse (a partire da quello x) e l'accelerazione totale. Dai grafici emerge che la relazione tra queste due variabili è molto stabile, il che suggerisce che l'accelerazione totale non dipende dall'attività in corso. Inoltre, sono presenti valori speculari che indicano una relazione simile a quella sinusoidale. Le stesse considerazioni valgono anche per la correlazione tra l'accelerazione lungo l'asse y e l'accelerazione totale corrispondente.

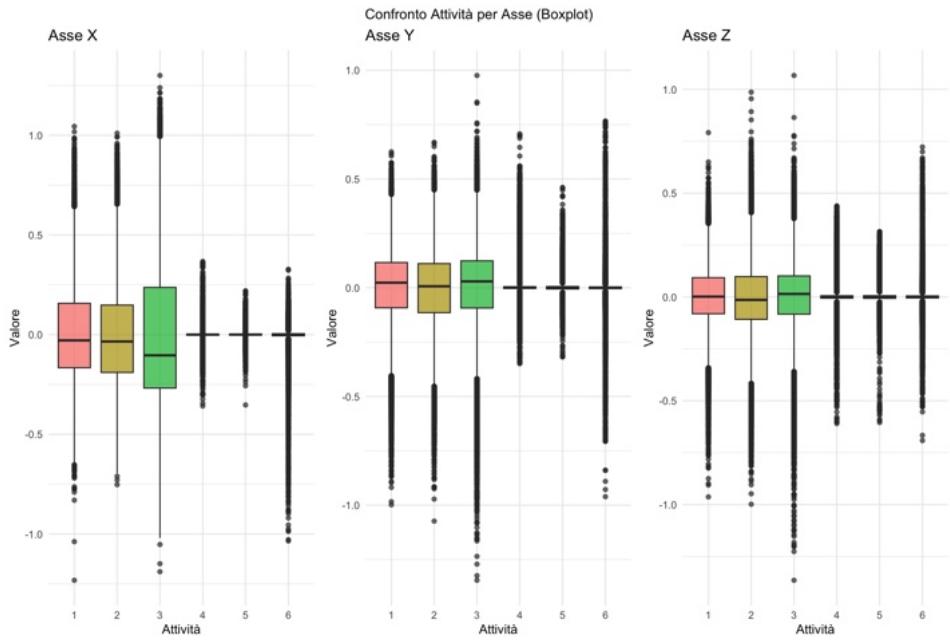


Figure 13: Boxplot di Attività per ogni Asse - Accelerazione

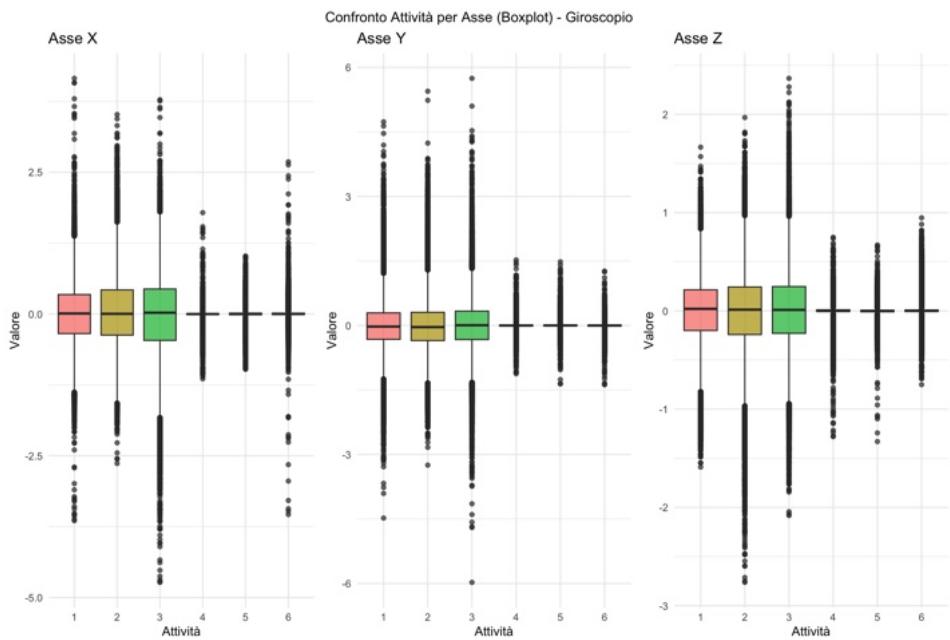


Figure 14: Boxplot di Attività per ogni Asse - Giroscopio

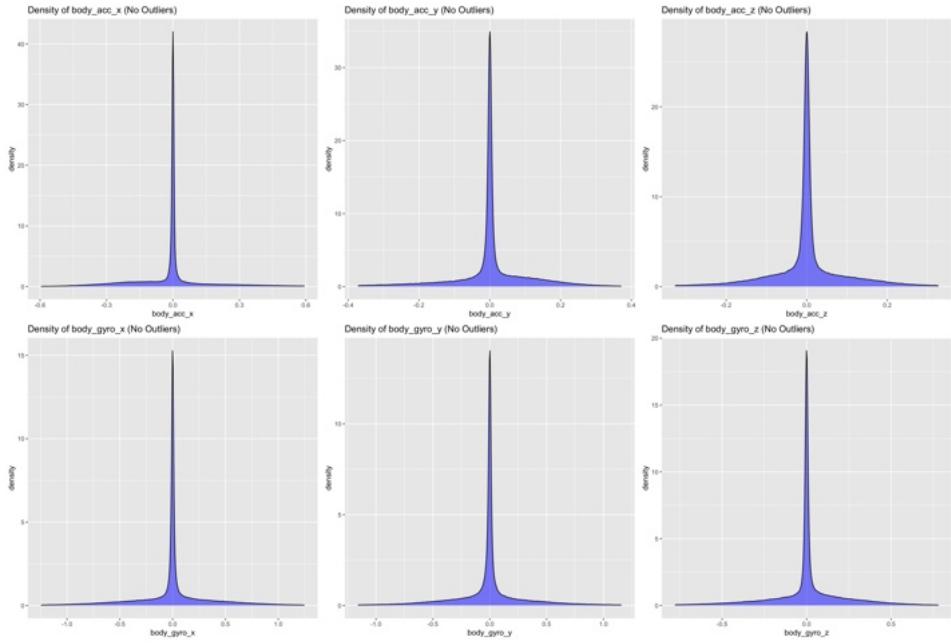


Figure 15: Distribuzione delle variabili senza outliers.

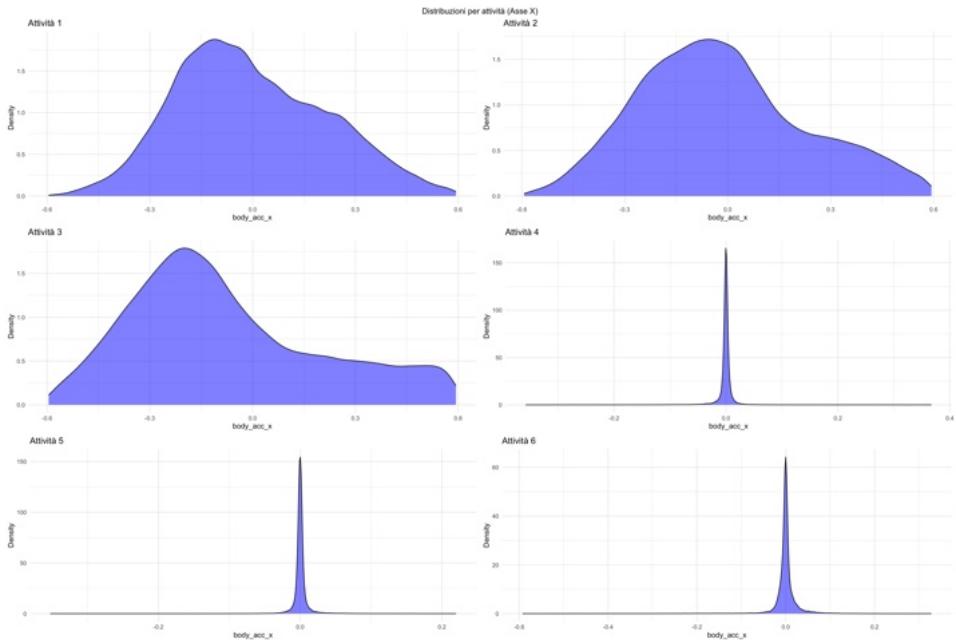


Figure 16: Distribuzione della variabile x senza outliers tra le varie attività.

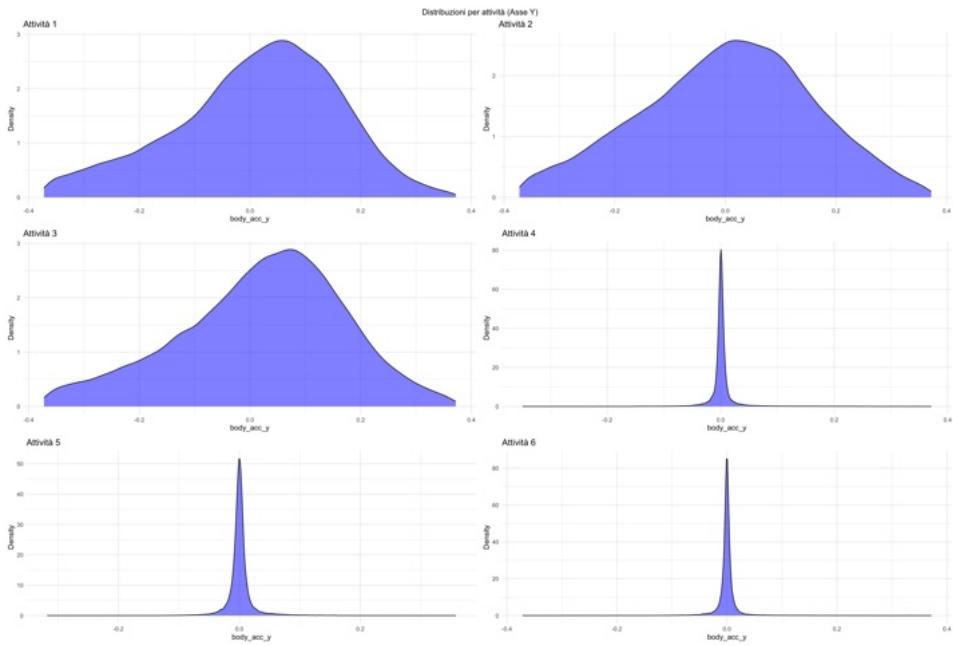


Figure 17: Distribuzione della variabile y senza outliers tra le varie attività.

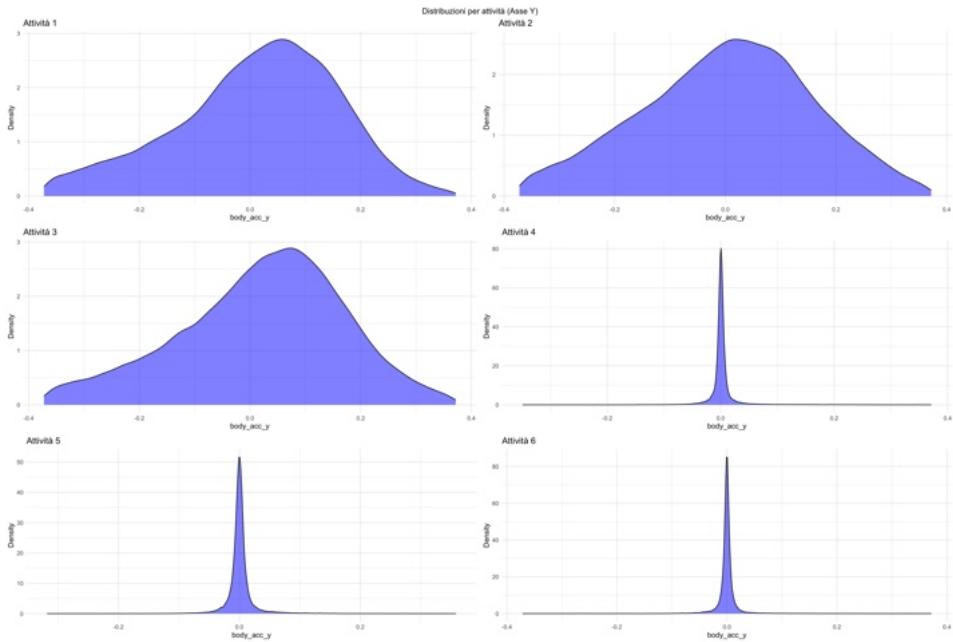


Figure 18: Distribuzione della variabile z senza outliers tra le varie attività.

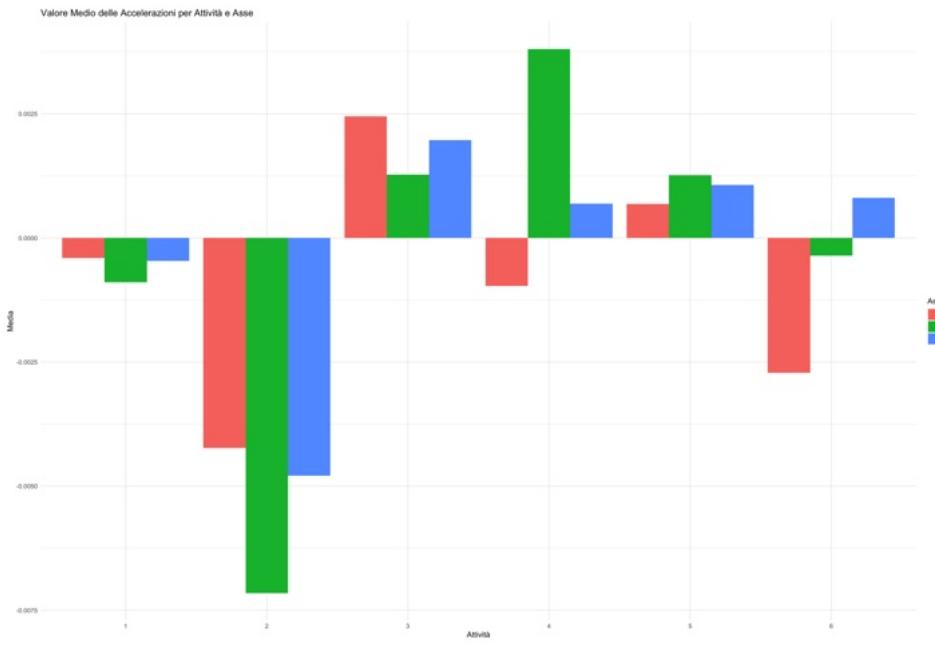


Figure 19: Valore medio delle accelerazioni per attività e asse.

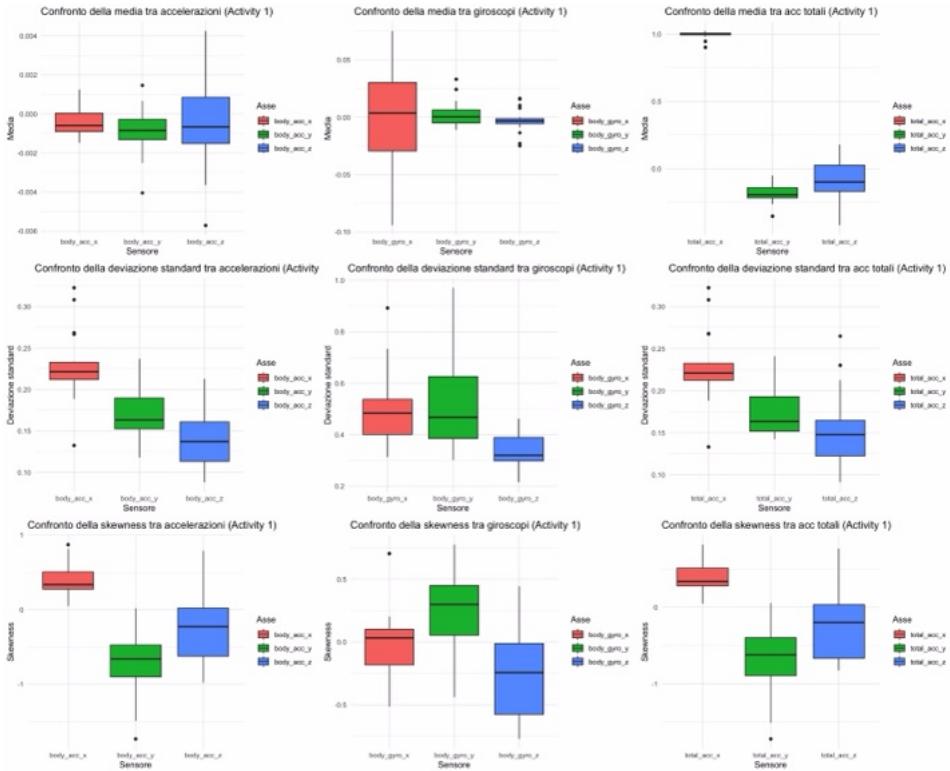


Figure 20: Confronto metriche attività camminata

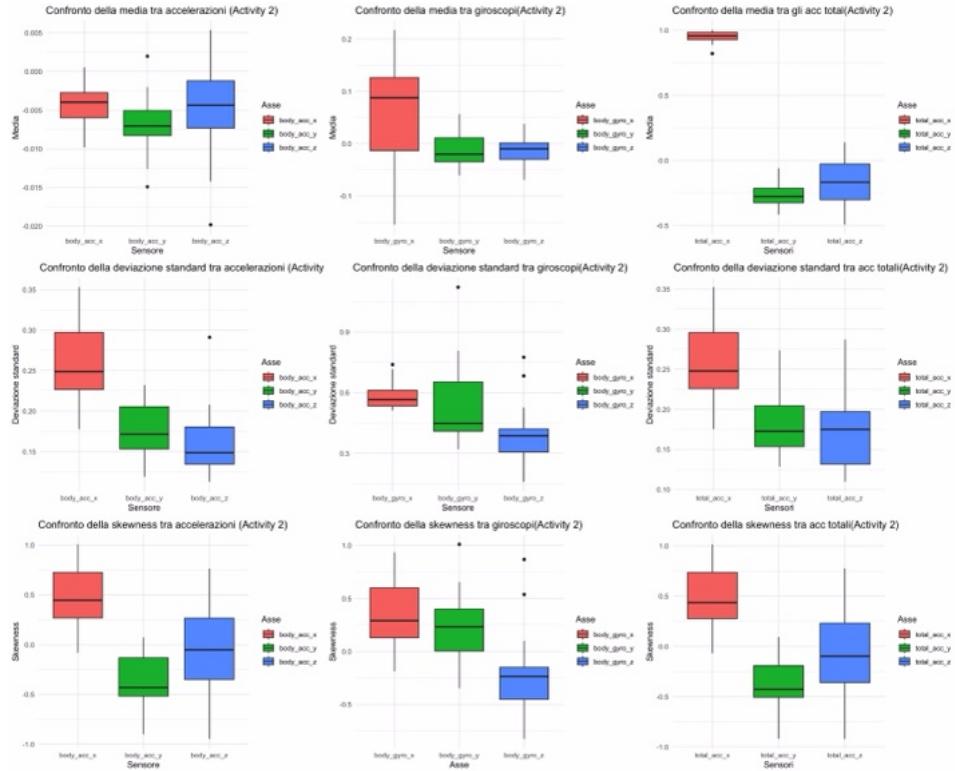


Figure 21: Confronto metriche attività camminata a salire

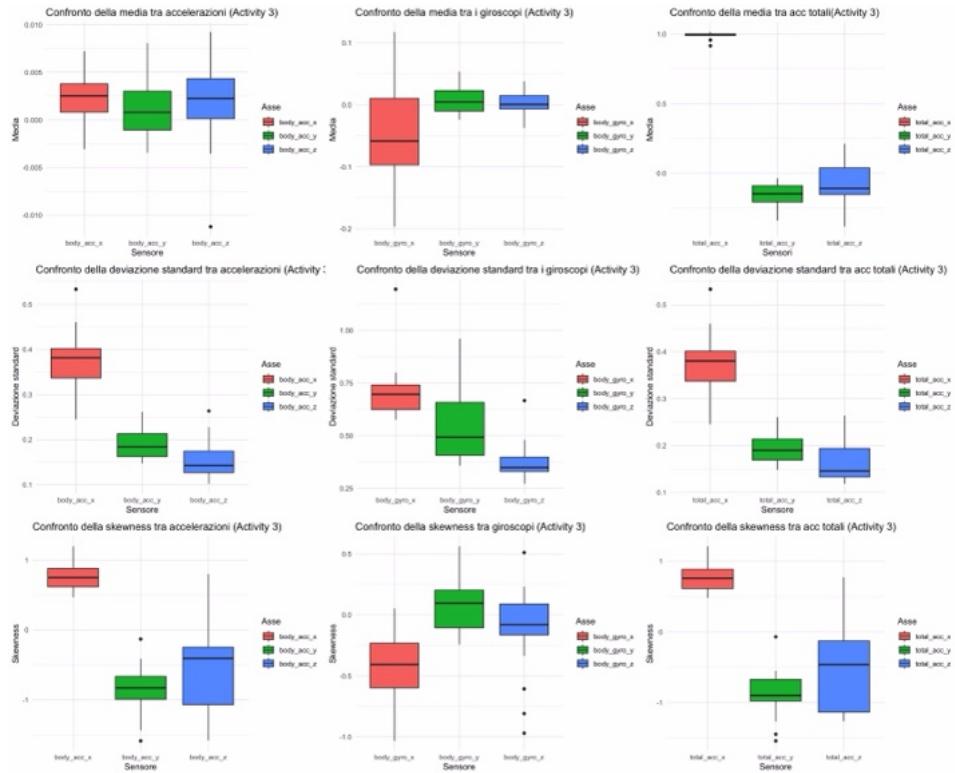


Figure 22: Confronto metriche attività camminata a scendere

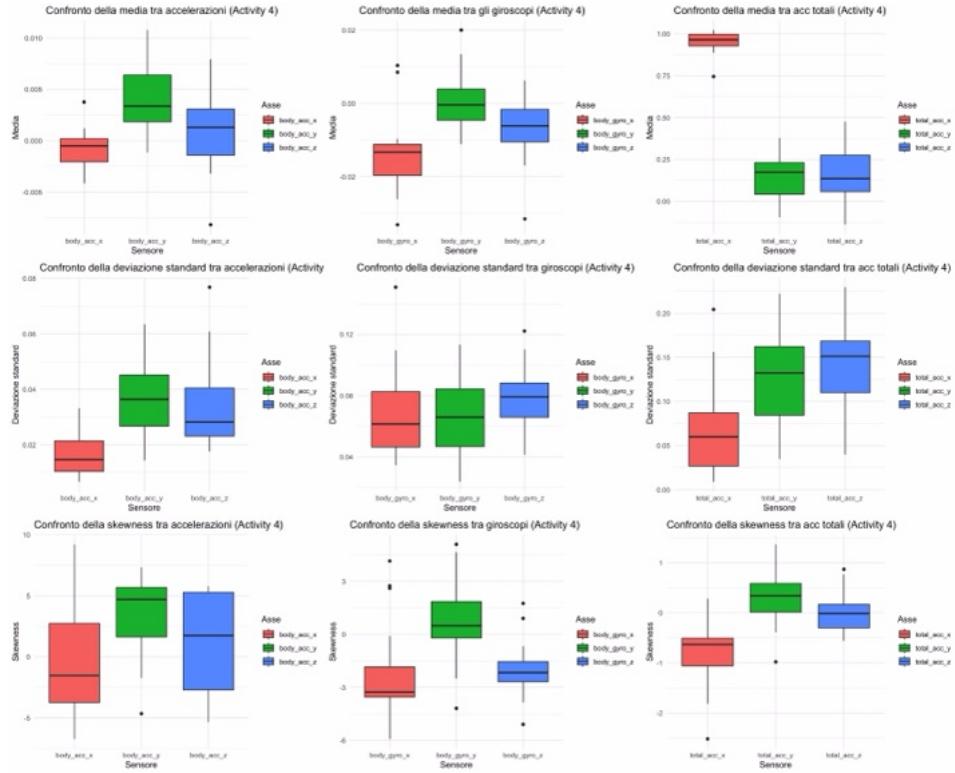


Figure 23: Confronto metriche attività seduto

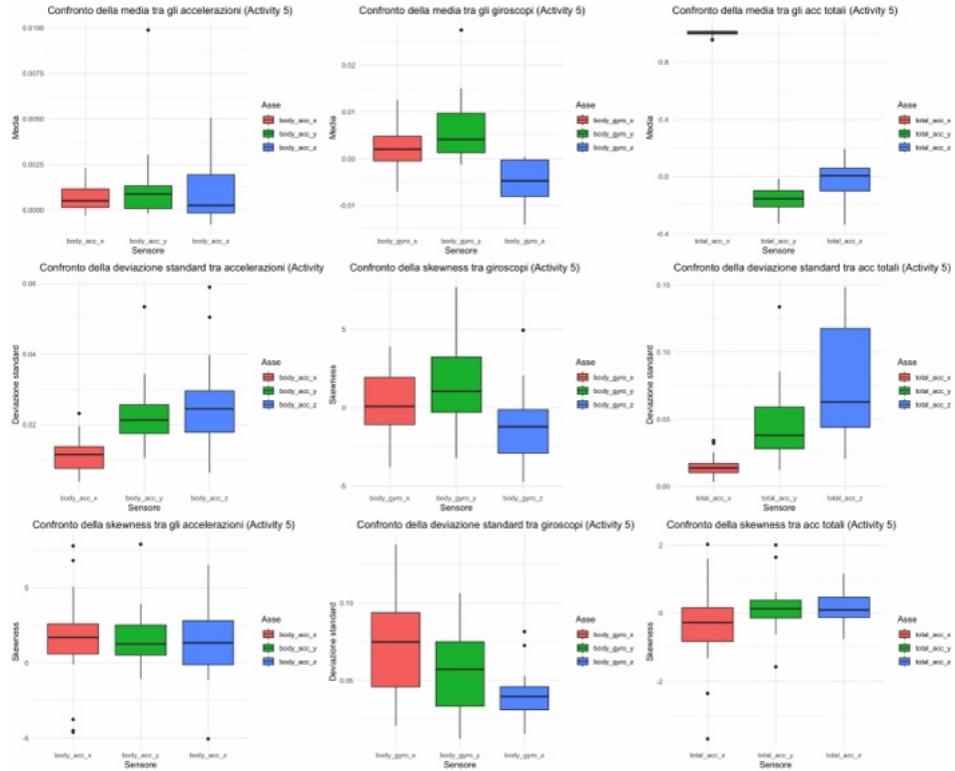


Figure 24: Confronto metriche attività in piedi

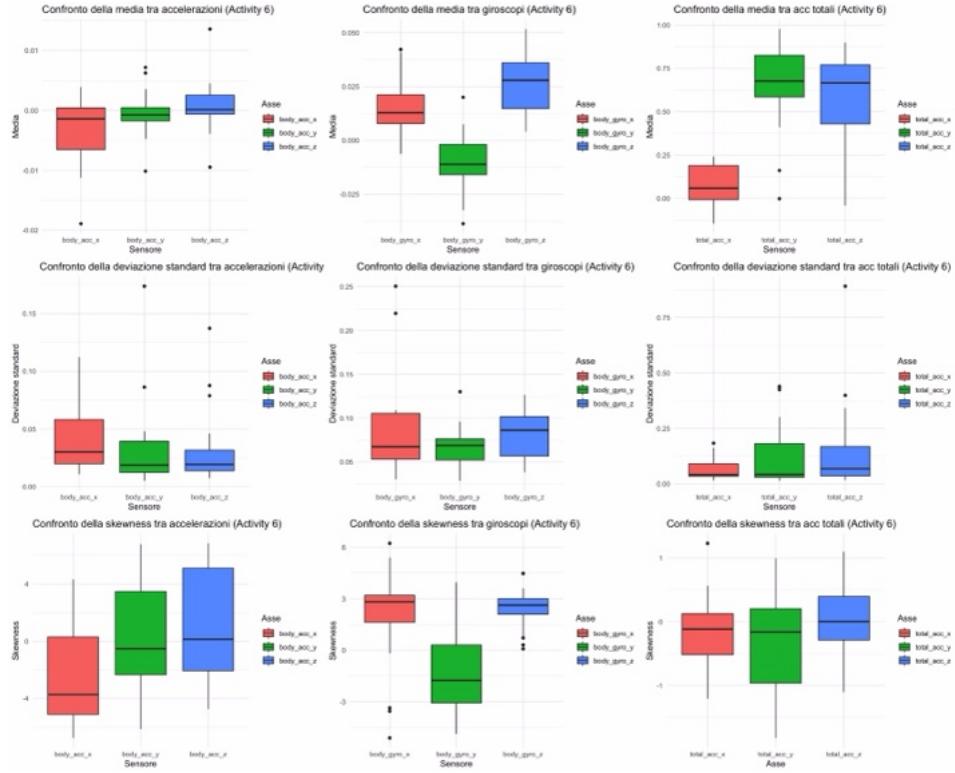


Figure 25: Confronto metriche attività sdraiato

Un'altra caratteristica che si evince è la presenza di picchi sia a lag negativi sia a lag positivi, evidenziando così una relazione bidirezionale che varia nel tempo.

L'ultima serie di grafici mostra la correlazione tra le varie accelerazioni totali, a partire dalla correlazione tra l'asse x e l'asse y. Dai grafici notiamo che inizialmente la correlazione negativa, fino alla quinta finestra è completamente negativa, dalla quinta in poi presenta una situazione alternata, inoltre i grafici differiscono tra loro, suggerendo che la relazione è influenzata dall'attività. Lo stesso ragionamento vale per la correlazione tra l'accelerazione totale dell'asse x con l'accelerazione totale dell'asse z, solo che dalla quinta in poi si stabilizza su una correlazione prevalentemente positiva nelle successive finestre per poi alternarsi in negativo e positivo. La correlazione tra l'accelerazione totale delle y e delle z invece, si comporta al contrario, inizialmente c'è una correlazione prevalentemente positiva, successivamente inizia ad alternarsi con leggeri picchi di correlazione negativa per poi alternarsi sempre di più, anche in questo caso la relazione dipende dal tipo di attività che si sta svolgendo.

### 3.3 Serie Temporali

In seguito all'analisi bivariata, abbiamo proseguito lo studio applicando la decomposizione STL (Fig. 34) per verificare se una determinata misurazione presenti trend o stagionalità differenti a seconda dell'attività e per comprendere il comportamento della variabile durante l'esecuzione dell'attività. Se il trend varia tra le diverse attività, ciò indica che la variabile è significativa per distinguerle. Analogamente, una stagionalità caratterizzata da oscillazioni

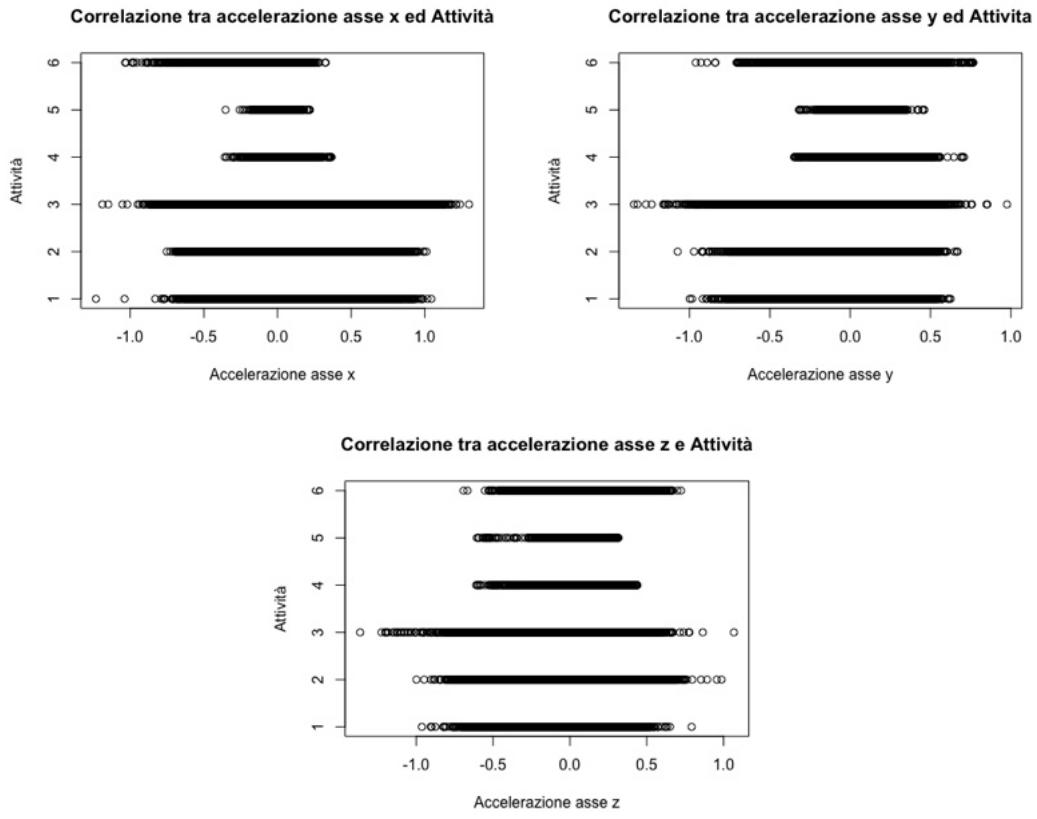


Figure 26: Scatterplot accelerazioni - Attività

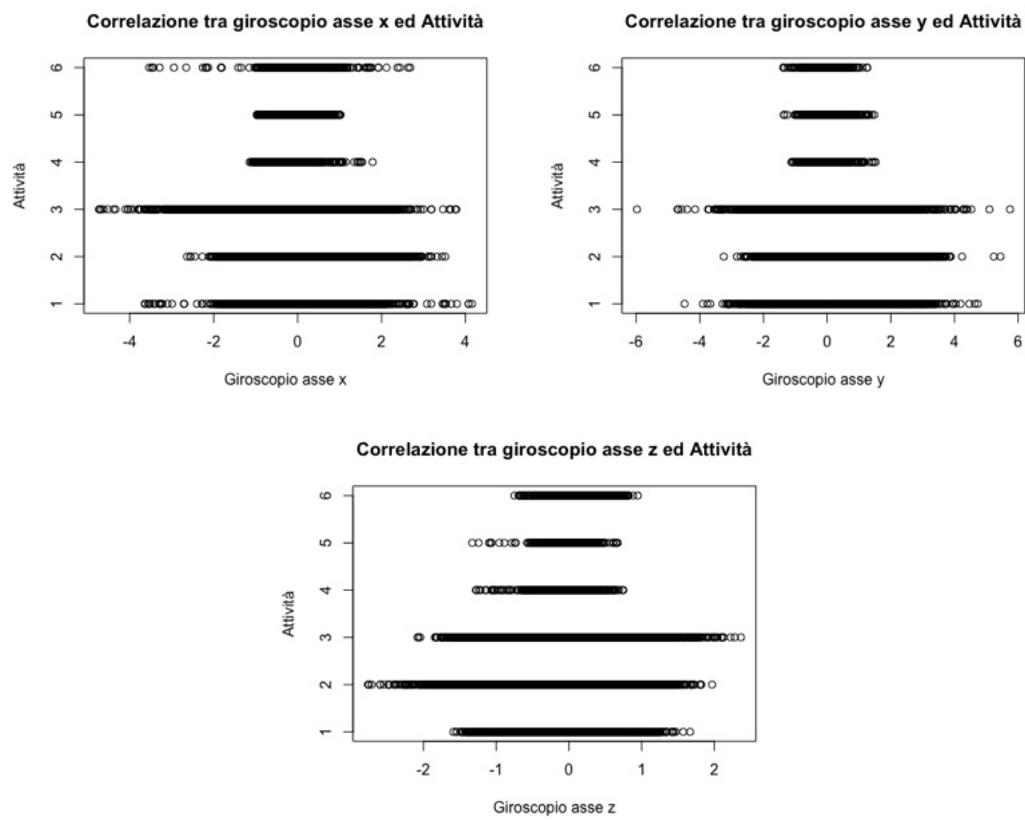


Figure 27: Scatterplot giroscopi - Attività

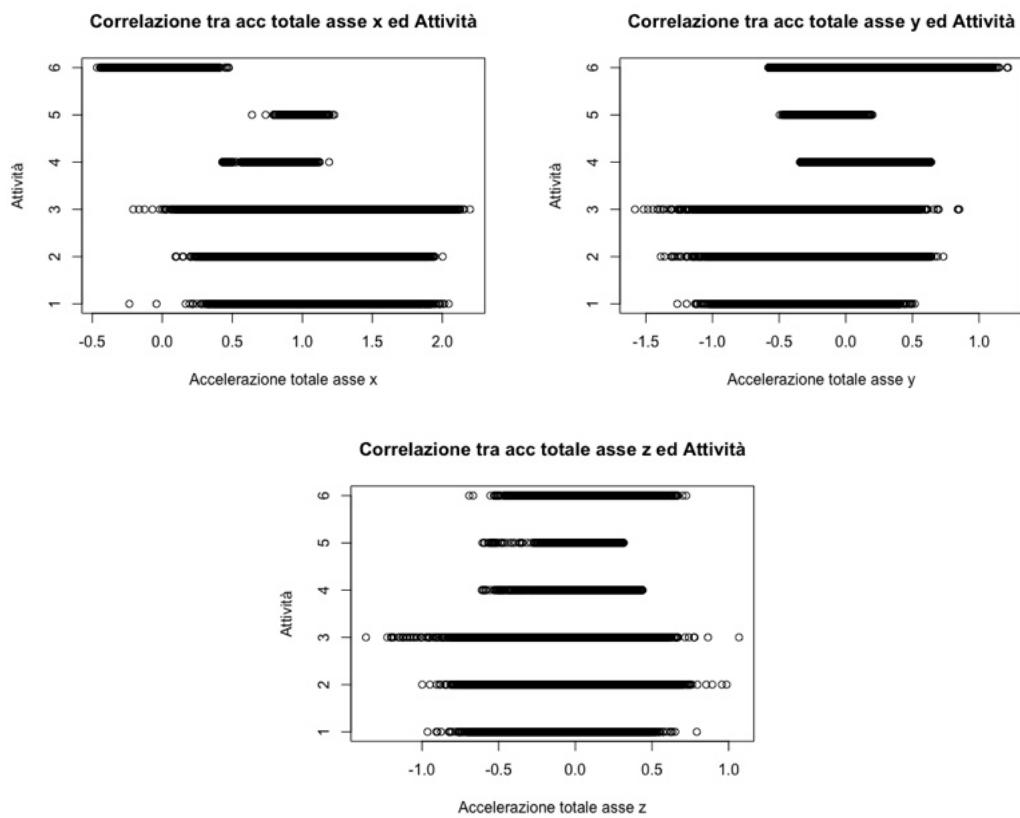


Figure 28: Scatterplot accelerazione totale - Attività

marcate e frequenti suggerisce che l'attività è accompagnata da movimenti ripetitivi e bruschi.

Per tutte le variabili analizzate, sia il trend che la stagionalità differiscono, il che evidenzia la loro utilità nella classificazione dell'attività, uno dei nostri obiettivi principali.

Dopo questa panoramica generale, osserviamo il comportamento di ciascuna variabile all'interno delle diverse attività. Esaminando i grafici dell'accelerazione sull'asse x, notiamo che nelle prime tre attività, che rappresentano i vari tipi di camminata, l'accelerazione oscilla ciclicamente tra valori positivi e negativi. Nelle ultime tre attività, invece, l'accelerazione risulta più stabile e ripetitiva. Ad esempio, nella quinta attività, che corrisponde allo stare fermi, si osserva un'accelerazione inizialmente positiva (probabilmente dovuta al movimento iniziale del soggetto), che successivamente si azzera e poi mostra un lieve incremento dovuto a piccoli movimenti.

La dinamica dell'accelerazione lungo l'asse y è invece diversa: sebbene nelle prime tre attività si osservino delle oscillazioni, nella quinta attività il sensore registra comunque delle accelerazioni, probabilmente a causa di leggeri movimenti del soggetto. Per quanto riguarda l'accelerazione sull'asse z, in ogni attività si osserva un'iniziale forte accelerazione che poi tende a stabilizzarsi.

Gli altri grafici analizzano le serie temporali relative ai giroscopi, a partire da quello sull'asse x.

## 4 Spiegazione Metriche e Metodologie

Prima di presentare i risultati ottenuti dalla nostra analisi, riteniamo opportuno introdurre e spiegare le metriche di valutazione e le metodologie statistiche impiegate. Questo permetterà una migliore comprensione dei valori riportati e delle conclusioni tratte.

### 4.0.1 Adjusted Rand Index (ARI)

L'Adjusted Rand Index (ARI) è una metrica cruciale per valutare la similarità tra due partizioni di dati, specialmente quando si confronta una clusterizzazione ottenuta tramite K-Means sui dati non etichettati con una classificazione con etichette reali (nel nostro caso le attività o i pazienti). L'ARI misura quanto due assegnazioni di cluster sono simili, correggendo per la similarità che potrebbe emergere puramente per caso.

Nel contesto della nostra analisi, abbiamo utilizzato l'ARI per determinare se i cluster identificati nei dati generati, senza utilizzare le etichette di attività o paziente, corrispondono effettivamente alle categorie reali (attività o pazienti). Un valore di ARI pari a 1 indicherebbe una perfetta corrispondenza tra i cluster trovati e le etichette reali. Un valore di 0 suggerisce che la similarità è paragonabile a quella che ci si aspetterebbe da assegnazioni casuali, indicando che i cluster non riflettono le categorie reali. Valori negativi segnalano una similarità inferiore al caso.

Formalmente, l'ARI è definito come:

$$ARI = \frac{RI - ExpectedRI}{MaxRI - ExpectedRI}$$

dove  $RI$  è il Rand Index,  $ExpectedRI$  è il Rand Index atteso sotto il modello nullo di

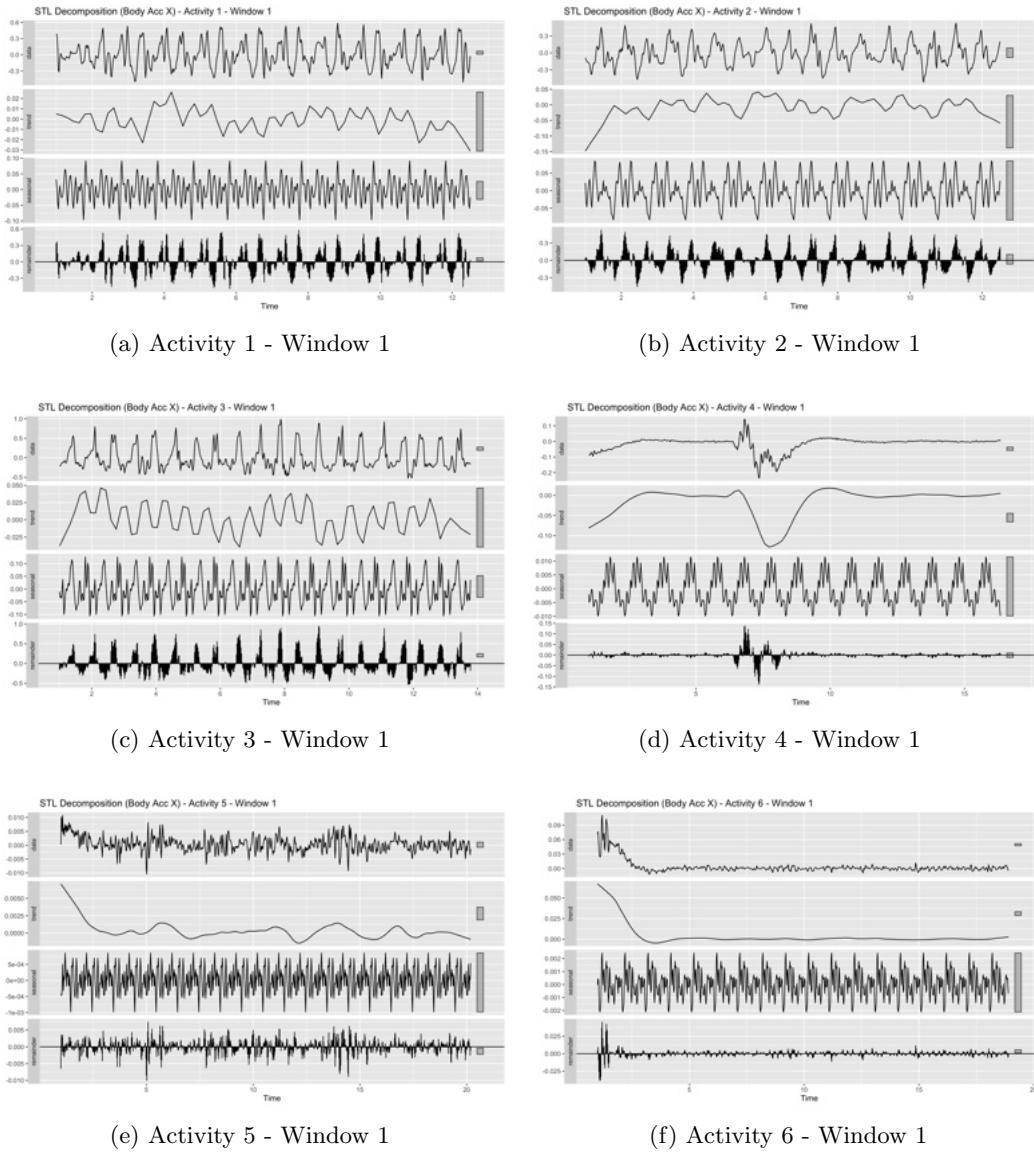


Figure 29: body\_acc\_x - Decomposizione STL delle Attività per il Subject 1

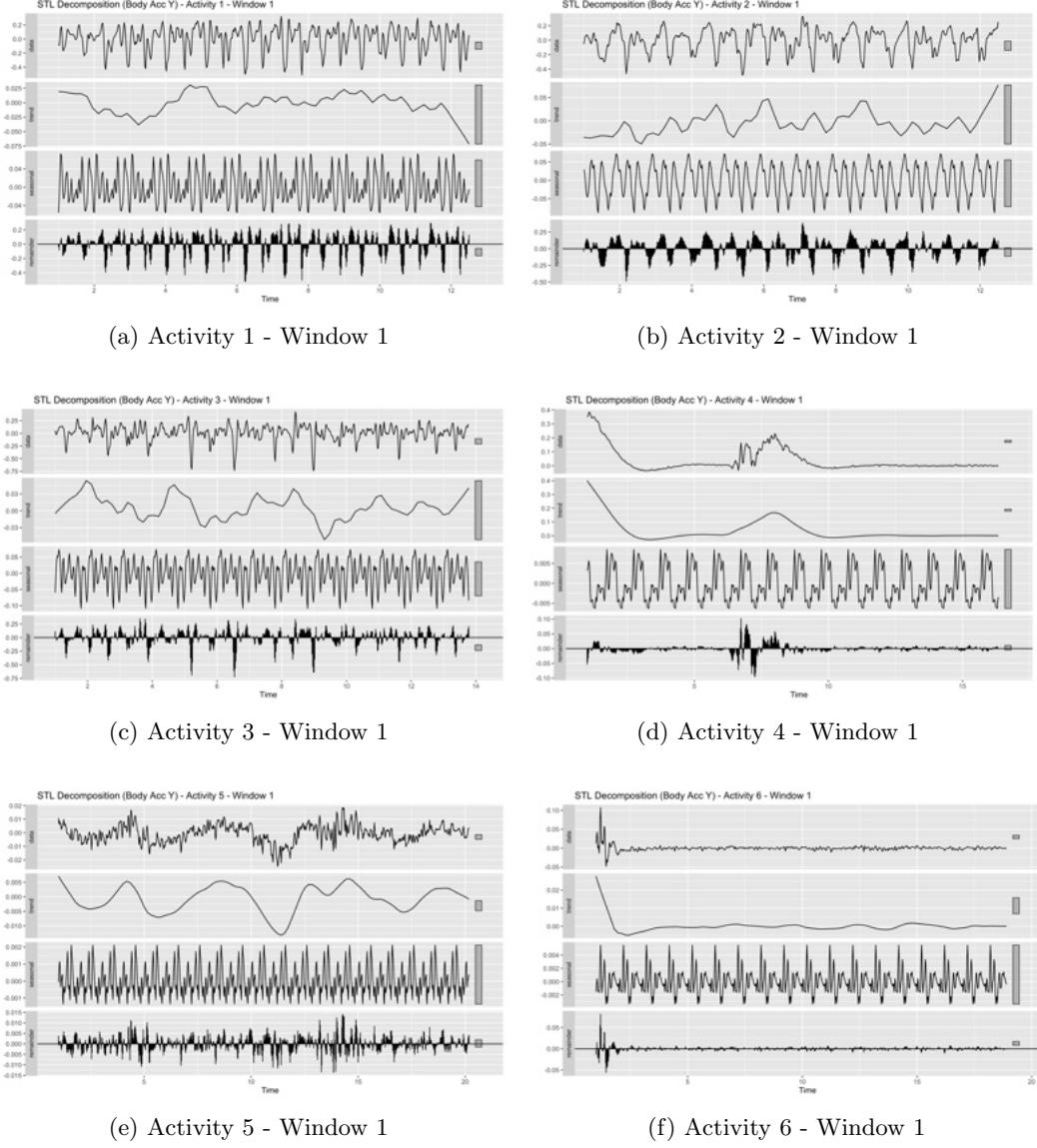


Figure 30: body\_acc\_y - Decomposizione delle Attività per il Subject 1

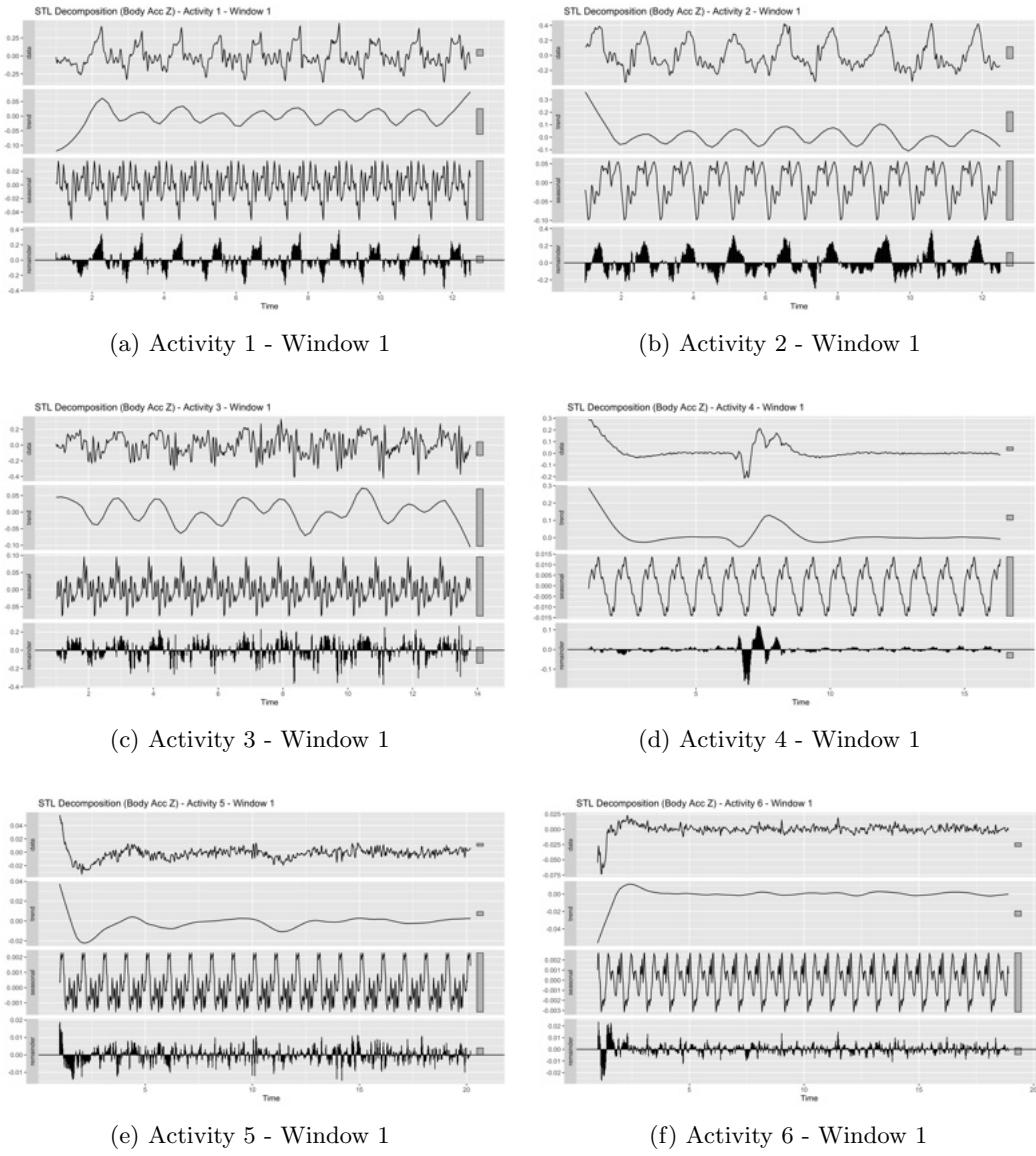


Figure 31: body\_acc\_z - Decomposizione STL delle Attività per il Subject 1

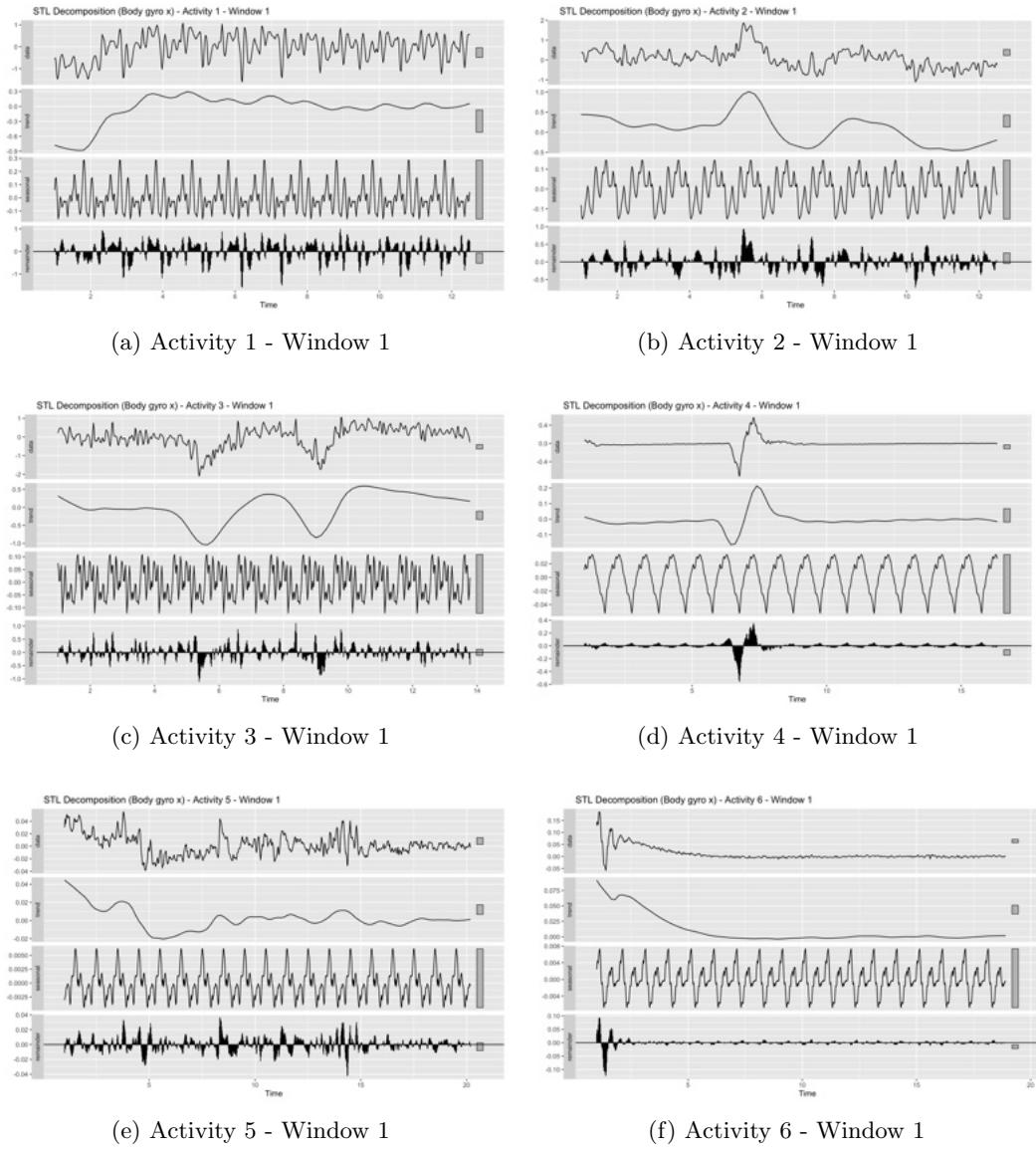


Figure 32: body\_gyro\_x - Decomposizione STL delle Attività per il Subject 1

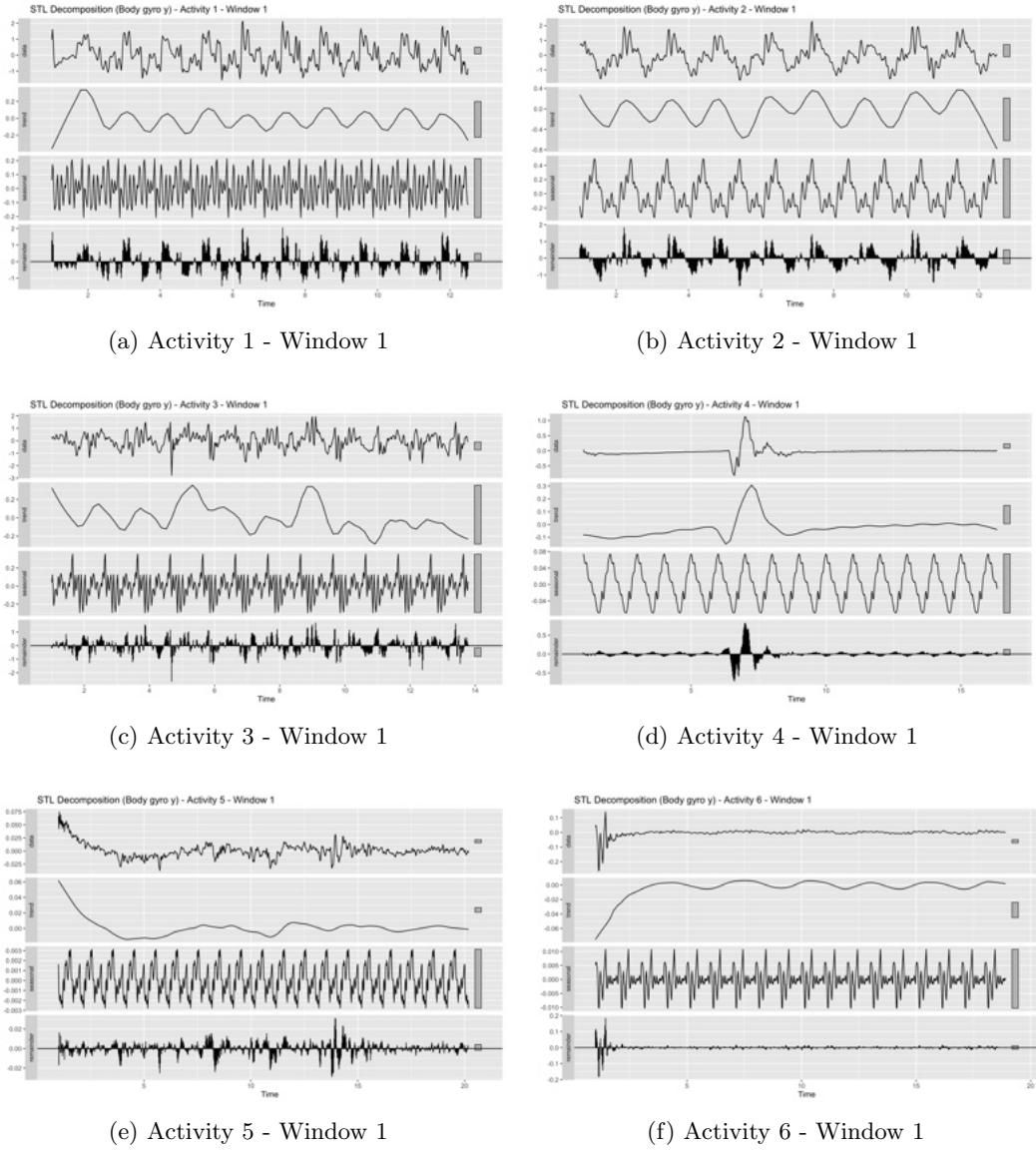


Figure 33: body\_gyro\_y - Decomposizione STL delle Attività per il Subject 1

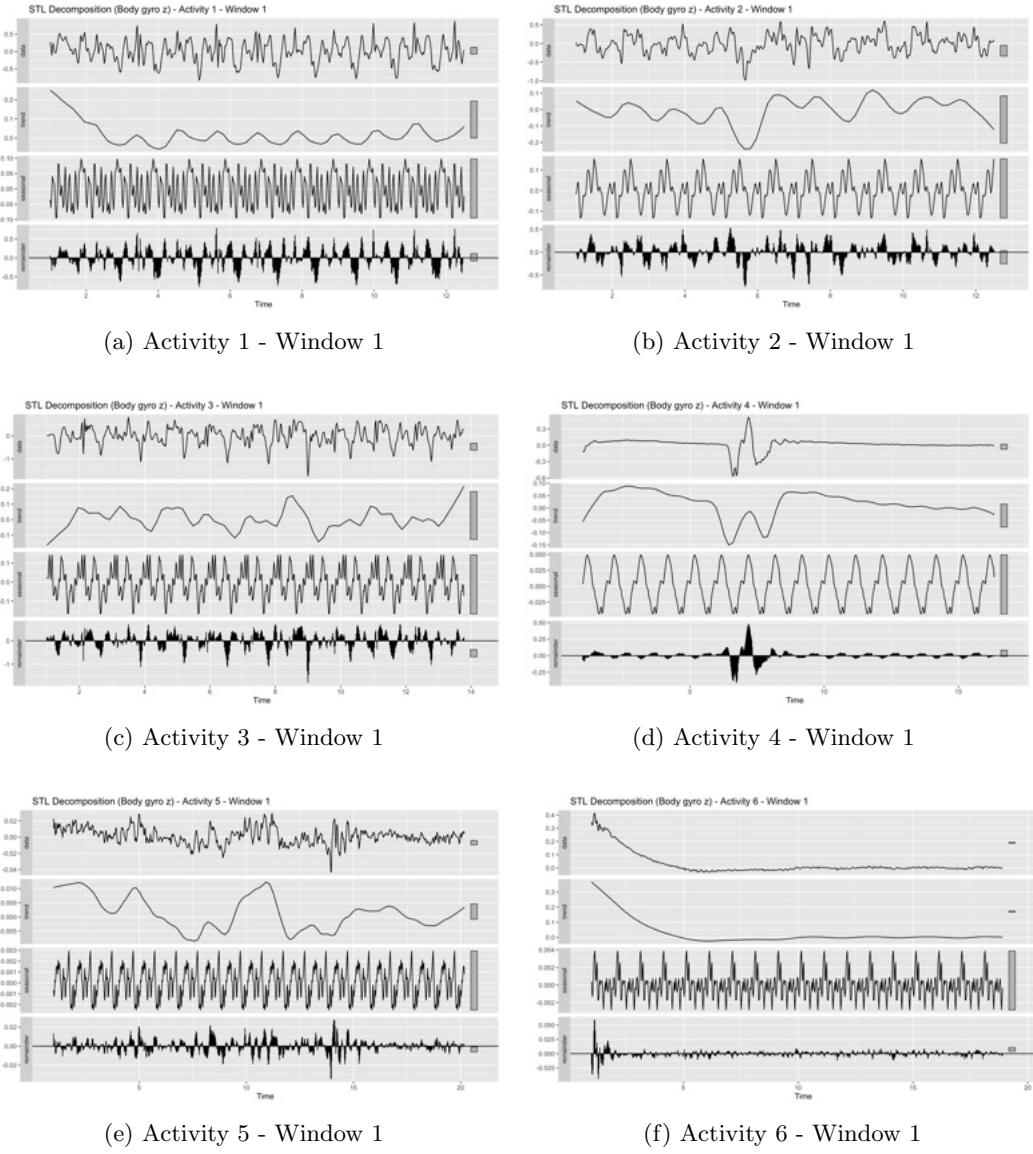


Figure 34: body\_gyro\_z - Decomposizione STL delle Attività per il Subject 1

assegnazione casuale, e *MaxRI* è il valore massimo del Rand Index.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

#### 4.0.2 V-Measure

La V-Measure è un'altra metrica importante che abbiamo impiegato per valutare la qualità della clusterizzazione in relazione alle etichette reali. Essa valuta la clusterizzazione considerando sia l'omogeneità che la completezza rispetto alle classi vere (etichette).

Nel nostro caso, la V-Measure ci aiuta a capire quanto bene i cluster ottenuti dai dati generati, senza conoscere le etichette, riflettono le categorie di attività o pazienti.

- **Omogeneità:** In questo contesto, l'omogeneità misura quanto i cluster identificati contengono esclusivamente finestre di dati che appartengono alla stessa etichetta reale (ad esempio, alla stessa attività o allo stesso paziente). Un cluster omogeneo contiene idealmente solo dati di una singola classe.
- **Completezza:** La completezza misura quanto tutte le finestre di dati che appartengono a una determinata etichetta reale (attività o paziente) sono state assegnate allo stesso cluster identificato. Idealmente, tutti i dati di una classe dovrebbero essere raggruppati in un unico cluster.

La V-Measure è la media armonica tra omogeneità ( $h$ ) e completezza ( $c$ ):

$$v = \frac{(1 + \beta) \times h \times c}{\beta \times h + c}$$

Solitamente,  $\beta$  è impostato a 1, dando uguale peso a omogeneità e completezza. In tal caso, la formula si semplifica a:

$$v = \frac{2 \times h \times c}{h + c}$$

I valori di V-Measure variano tra 0 e 1, dove 1 indica una clusterizzazione perfetta rispetto alle etichette reali, e 0 indica una clusterizzazione che non riflette in alcun modo le etichette reali. Valori bassi di V-Measure, come quelli ottenuti nella nostra analisi, suggeriscono che i cluster trovati non corrispondono alle categorie reali (attività o pazienti).

## 5 Risposta alla Prima domanda di Ricerca

**È possibile determinare un'attività da una finestra di misurazioni consecutive senza conoscere il paziente?** Questo è equivalente a provare la seguente ipotesi: "Le distribuzioni multivariate delle feature (accelerazioni lineari e velocità angolari) differiscono significativamente tra attività."

La prima cosa che abbiamo provato è stato il clustering Kmeans per creare clustering delle singole istanze. Ma ciò non ha funzionato e non si riuscivano a creare cluster significativi che distinguessero le varie attività e quindi considerare le misurazioni di un singolo istante, senza tener conto del contesto in cui è inserita la finestra, si è rivelato logicamente poco utile. Così abbiamo creato activity\_features, che contiene un sommario delle attività

(raggruppate per utente, attività e finestra temporale), ma ancora una volta, anche se comunque questa volta le metriche erano migliori, rimanevano comunque molto basse (ARI 0.22, V-Measure 0.3). A questo punto abbiamo pensato di procedere per passi: Innanzitutto abbiamo ridotto le 6 attività a 2 per rendere la distinzione più evidente: Da una parte le Attività che richiedono movimento e dall’altra le attività che non lo richiedono (Le attività che richiedono movimento – ovvero camminata, salire le scale e scendere le scale – sono state codificate con il valore 1, mentre quelle statiche – stare fermi e stare in piedi – con il valore 0.). Partendo con questa divisione, era chiaro che, se il modello non fosse neanche in grado di distinguere un’attività in movimento da una statica allora non sarebbe stato necessario continuare. Tuttavia, come è visibile anche nell’immagine [35], il KMeans è in grado di distinguere due cluster che corrispondono proprio alle due attività che abbiamo creato, con un ARI e V-measure maggiori del 96%.

A questo punto, abbiamo considerato il sottoinsieme di tutte e solo le attività che coinvolgono il movimento, ma il modello di clustering in questo caso non riusciva a ottenere cluster significativi che effettivamente corrispondessero a varie etichette. Prima di arrendersi, però abbiamo analizzato il grafico di due componenti e anche se le componenti erano molto vicine, è possibile notare che banalmente più alta è la PC2, più probabile è che il singolo elemento corrisponda all’attività di salire le scale, mentre più bassa è, più probabile è che sia scendere le scale. Dunque abbiamo rimosso l’attività di walking e abbiamo provato la regressione lineare su queste due PC per prevedere la terza ed effettivamente le metriche sono risultate abbastanza alte. Ora, non ci restava che analizzare le attività che non richiedono movimento: Stare in Piedi; Stare Seduti; Sdraiati. Effettuando il cluster, di nuovo non riusciamo a distinguere per attività, ma osservando il grafico si potrebbe tracciare una linea che divide le attività stare sdraiati e stare in piedi: ciò viene confermato anche dalla regressione lineare [37b].

**Risposta alla domanda:** Concludendo, possiamo affermare che è possibile determinare se una specifica finestra temporale di Soggetto, Attività corrisponde ad un’Attività che comporta movimento (1,2,3) oppure che non lo comporti(4,5,6). Non riusciamo invece a distinguere esattamente a quale delle 3 corrispondano ma, escludendo l’attività 1, la camminata, possiamo distinguere tra camminata su per le scale e quella giù per le scale, ed escludendo lo stare seduti, possiamo distinguere lo stare in piedi dallo stare sdraiati.

## 6 Risposta alla Seconda Domanda di Ricerca

**È possibile determinare il paziente da una finestra di dati conoscendo l’attività?** Per il secondo obiettivo, abbiamo rimosso l’etichetta Subject e suddiviso il dataset in 6 parti, una per ciascuna attività. Successivamente, per ogni parte abbiamo creato un cluster per rispondere alla seguente domanda: Dato un insieme di finestre di una determinata attività, è possibile risalire al Subject di quella finestra? Infatti, se non siamo in grado di distinguere i pazienti conoscendo l’attività, non potremo farlo in assenza di tale informazione. Dai risultati del clustering [38] si evince che per le prime 3 attività (quelle in movimento), la correlazione tra cluster e Subject è bassa ma non trascurabile: per le prime due, l’ARI e la V-measure sono intorno allo 0.2 e 0,5, mentre per la prima attività (walking), sono 0,5 e 0,7. Invece per le ultime 3 attività, quelle da fermi, come logicamente ci aspettavamo, ARI e V-Measure sono <0.1. A questo punto testiamo l’ipotesi nulla, e cioè: “L’allineamento

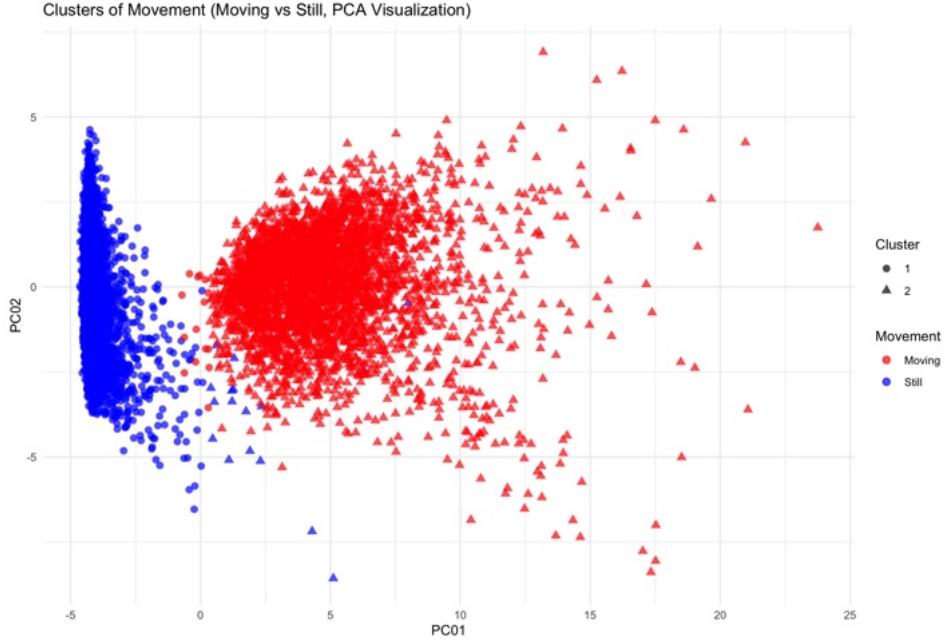


Figure 35: Movimento (1,2,3) vs Stare fermi (4,5,6) - Clustering

tra i cluster ottenuti e i soggetti è statisticamente migliore di quanto ci si aspetterebbe per caso?" Per testare l'ipotesi nulla abbiamo permutato le etichette dei pazienti: abbiamo preso le etichette vere e assegnate in modo random ai dati: ciò azzera ogni relazione tra i cluster e i soggetti. Dopo aver fatto questo processo diverse volte, ecco la media delle metriche ARI e V-Measure: 0.000 e 0.000, e entrambe con un p-value di 0,01. Il test, con p-value <0,05, permette di rifiutare l'ipotesi nulla, confermando così una significativa correlazione tra i cluster e i soggetti. Dunque, anche se non si può riconoscere con esattezza a quale subject corrisponde (ARI e V-Measure di 0,01 senza separazione di attività), c'è una forte correlazione tra ogni finestra temporale e lo specifico Subject. **Risposta alla domanda:** Conoscendo l'attività, il metodo permette di determinare il soggetto in maniera parziale. In particolare, per le attività dinamiche (ad esempio, walking) il clustering mostra una discreta capacità discriminante (ARI fino a 0,5 e V-measure fino a 0,7), mentre per le attività statiche le performance sono molto basse (<0,1).

## 7 Analisi di un Dataset Generato

Da questa sezione in poi, lo studio si è focalizzato su un dataset relativo al riconoscimento di attività generato da un Large Language Model. Nelle prime sezioni si effettueranno le stesse analisi effettuate nel dataset originale, confrontando anche i risultati. La parte finale invece si concentrerà sull'inferenza statistica, in particolare nel capire se il dataset generato si rifà a una distribuzione nota.

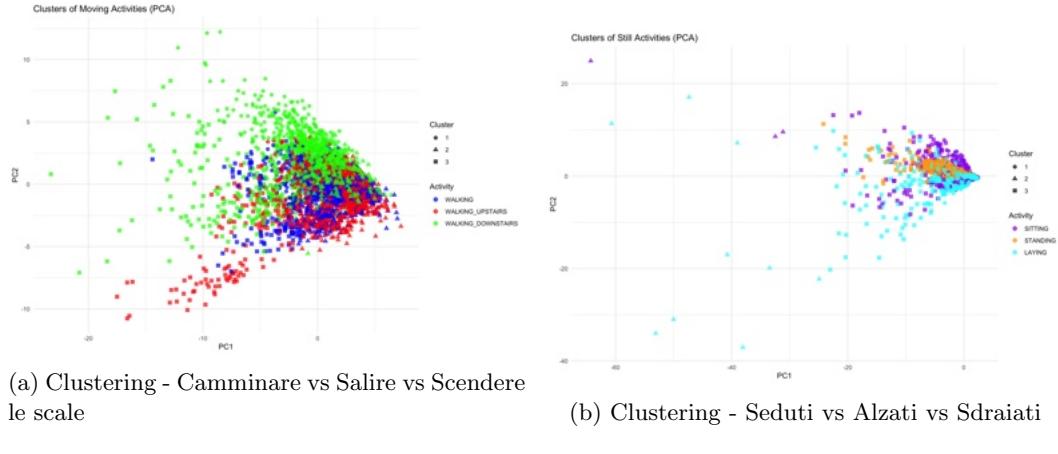


Figure 36: Clustering

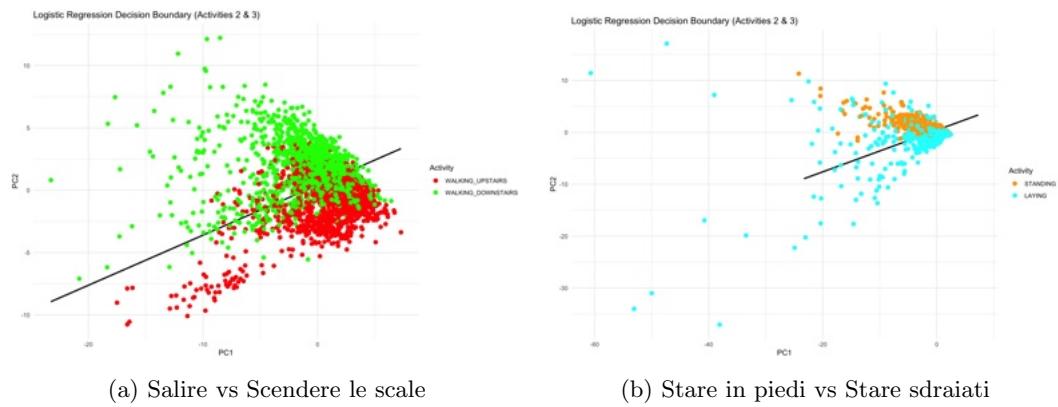


Figure 37: PCA con Regressione

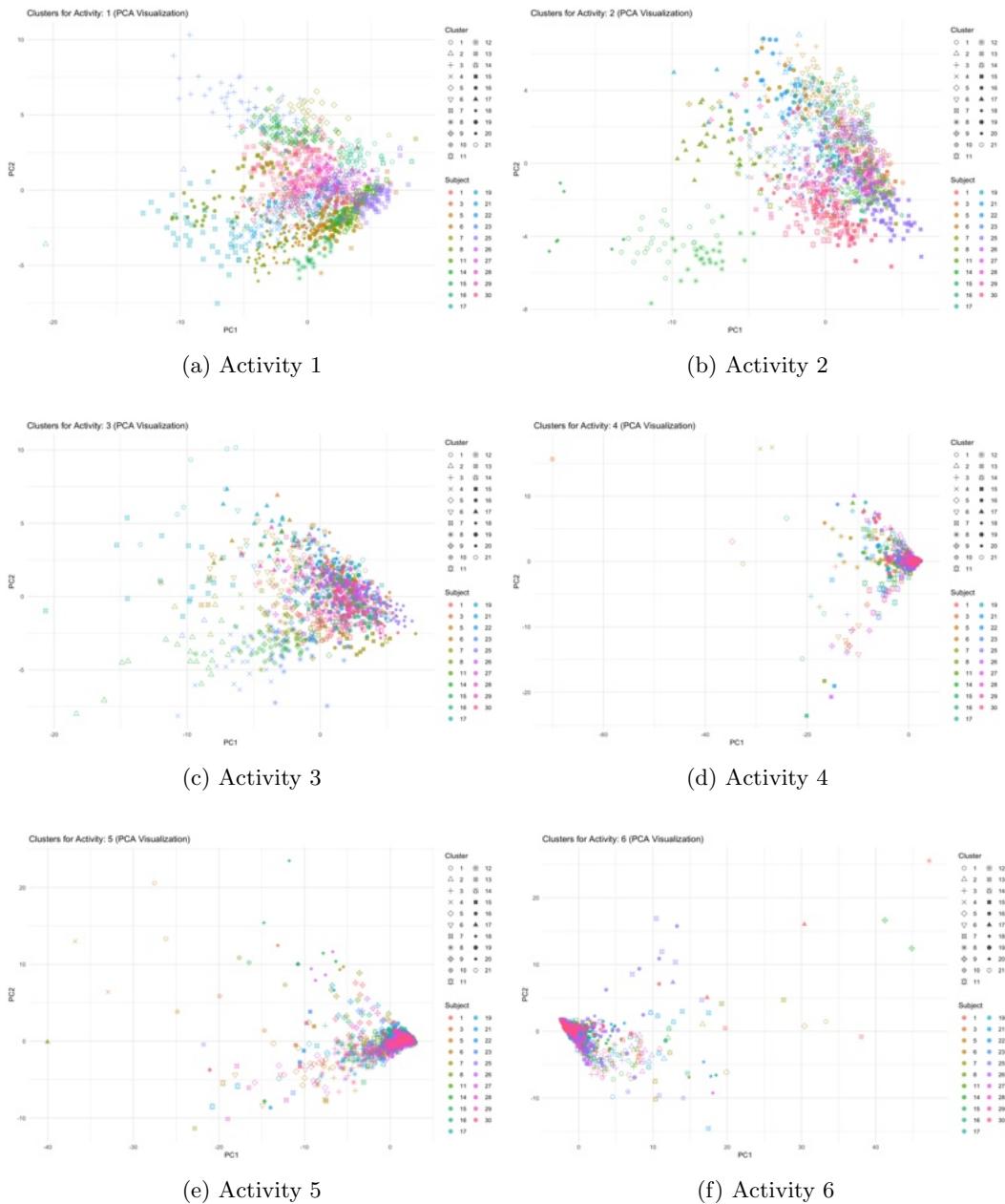


Figure 38: Clustering KMeans per la seconda ipotesi

## 7.1 Generazione del dataset

Il Large Language Model scelto per la generazione del dataset sintetico è stato ChatGPT, fornendogli un prompt quanto più accurato possibile, di seguito è mostrato il prompt che abbiamo fornito:

```
Agisci come un esperto di data synthesis per applicazioni di Human
Activity Recognition (HAR). Genera un dataset sintetico
multivariato che replichi le caratteristiche del dataset reale
descritto, nessun codice di alcun tipo, voglio un dataset, seguendo
rigorosamente queste specifiche:

Struttura Richiesta:
Colonne:

timestamp (incrementi di 20 ms, formato: 0.000, 0.020, 0.040, ...)
activity_label (valori esatti: "Walking", "Walking Upstairs", "Walking
Downstairs", "Sitting", "Standing", "Laying")

subject_id (ID unico da 1 a 30, con distribuzione casuale)

acc_x, acc_y, acc_z (accelerazione lineare in g, range 2g con rumore
Gaussiano =0.01-0.05)

gyro_x, gyro_y, gyro_z (velocità angolare in rad/s, range 4 con
rumore Gaussiano =0.02-0.1)

Pattern Attività:
Dynamic Activities (Walking/Upstairs/Downstairs):
Accelerazione: oscillazioni periodiche (frequenza 0.5-2 Hz per walking
, 1-3 Hz per stairs)

Giroscopio: picchi durante cambi di direzione (es. salita/discesa
scale)

Static Activities (Sitting/Standing/Laying):
Accelerazione: media 1g sull'asse verticale (es. asse Y se
smartphone verticale)

Giroscopio: valori quasi nulli (0 .1 rad/s), con micro-vibrazioni
casuali

Vincoli:
Frequenza di campionamento: 50 Hz (1 riga ogni 20 ms)
```

```

Finestre temporali: 2.56 sec      128 righe consecutive per finestra

Variabilit inter-soggetto: differenze fisiologiche (es. ampiezza
    segnale 15 % tra soggetti)

Istruzioni Aggiuntive:
Simula artefatti realistici (es. brevi spike nel giroscopio per
    movimenti bruschi)

Bilancia la distribuzione delle attivit ( 16 -17% per classe)

Includi una leggera deriva temporale nei segnali (es. offset crescente
    dopo 1 minuto di attivit )

Usa seed casuali riproducibili per soggetti specifici (es. subject_id
    = 5      stesso pattern se rieseguito)

Inoltre considera il fatto che sono presenti anche pi soggetti e non
    uno solo, quindi devi generare per soggetto, ti fornisco anche un
    esempio delle misurazioni:

```

Nel fornire un esempio delle misurazioni, si è fornito un campione casuale al modello. Da tale prompt ChatGPT è riuscito a crearcì una dataset di 30.000 istanze e 11 feature.

## 7.2 Struttura del Dataset

Come anticipato nella sezione precedente, il dataset generarato contiene 30.000 istanze e 11 feature, le feature sono le seguenti:

- Accelerazioni triassiali X, Y e Z;
- Velocità Angolare triassiale X, Y e Z;
- Attività che comprendono : Walking, Walking Upstairs, Walking Downstairs, Sitting, Standig e Laying;
- Soggetto di cui si effettua l'analisi in totale 10;
- Timestamp di misurazione lungo 10 secondi.

La prima cosa che si nota è che il dataset generato ha delle dimensioni molto inferiori rispetto a quello originale che era formato da circa 500 mila righe e analizzava 30 soggetti differenti. Inoltre il dataset generato non contiene le accelerazioni totali come in quello originale.

## 7.3 DataPreProcessing

Il dataset sintetico che ci ha fornito ChatGPT aveva delle incongruenze, la prima tra queste è che i valori delle accelerazioni, sia triassiali che angolari, non erano normalizzati in un range

di valori [-1, 1]. Per questo motivo è sembrato opportuno normalizzare i dati tramite la min-max normalization. Tale normalizzazione risulta necessaria poichè ci aiuta a interpretare meglio le cross-correlation tra le varibili prese in considerazione e facilita l'analisi delle serie temporali e delle decomposizioni.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times 2 - 1$$

Inoltre abbiamo scelto di cambiare la colonna TimeStamp (0,02;0,004;0,006) che incrementa ad una frequenza di 50hz arbitrariamente creata dall'LLM, e sostituito con un offset intero incrementale (1,2,3), per renderlo più facilmente utilizzabile per le misurazioni. In questo modo la Formattazione del Dataset è completamente uguale a quella del dataset originale dopo il preprocessing.

## 7.4 Analisi Univariata

**Come differiscono le metriche dei dati sintetici generati dagli Large Language Model dalle metriche dei dati reali?** Come per il dataset originale, abbiamo proseguito con un'analisi univariata delle feature del dataset, generando i kernel density plot, boxplot e istogrammi. Nella generazione di quest'ultimi abbiamo sia generato i grafici per l'intera feature, sia le metriche sintetiche. I grafici e le metriche sono stati calcolati sia in maniera generale, sia tenendo conto dell'attività che si stava svolgendo. L'analisi comparativa rivela differenze marcate nella struttura statistica:

- **Variabilità:** Le deviazioni standard (SD) nel dataset sintetico risultano sistematicamente più elevate per gli assi principali (0.734 vs 0.198 in body\_acc\_x), mentre si riducono drasticamente per quelli secondari (0.050 vs 0.260 in body\_gyro\_z). Questo pattern suggerisce una generazione non fisiologica dove il modello linguistico ha amplificato/diminuito arbitrariamente l'ampiezza dei segnali.
- **Distribuzioni:** I valori di curtosi nel sintetico sono prossimi allo zero (es. -1.293 vs 4.506 in body\_acc\_x), indicando distribuzioni iponormali, contrapposte alle marcate code pesanti dell'originale. I kernel density plot mostrano profili gaussiani artificiali nel sintetico vs distribuzioni multimodali legate alle attività fisiche nel reale.
- **Intervalli dinamici:** I range min-max del sintetico appaiono compressi e sottostimati in gyro (es.  $\pm 0.2$  vs  $\pm 2.5$  in body\_gyro\_z) nonostante SD maggiori, evidenziando una perdita degli outlier fisiologici tipici dei tracciati sensoriali.
- **Robustezza:** La Median Absolute Deviation (MAD) nel sintetico risulta inflazionata di un ordine di grandezza (1.001 vs 0.024 in body\_acc\_x), confermando la scarsa aderenza alle distribuzioni robuste tipiche dei dati reali.

Questo mismatch statistico si riflette nei visual analytics: **i boxplot del sintetico mostrano distribuzioni simmetriche e prive di skewness fisiologica**, mentre gli histogrammi dell'originale rivelano code asimmetriche legate alle transizioni tra attività. Particolarmente significativa è l'assenza nel sintetico delle distribuzioni leptocurtiche (kurtosis  $> 3$ ) caratteristiche dei segnali d'inerzia durante cambi di postura.

La Figura [40] evidenzia come nel dataset sintetico ogni attività mostri un profilo distributivo distintivo (es. picchi bimodali nella camminata, skew positivo nelle scale), mentre nell'originale tutte le attività condividono simili distribuzioni gaussiane centrate sullo zero, quasi come se l'LLM avesse immaginato delle firme cinematiche specifiche che nella realtà non esistono.

Table 5: Dataset Sintetico - Metriche riassuntive

Sensore	Media	Mediana	SD	Skewness	Kurtosis	Min	Max	MAD
body_acc_x	0.001	0.000	0.734	0.006	-1.293	-1.686	1.830	1.001
body_acc_y	-0.002	-0.005	0.735	-0.001	-1.290	-1.701	1.723	1.003
body_acc_z	-0.002	-0.002	0.200	0.003	-0.018	-0.770	0.755	0.201
body_gyro_x	0.000	0.003	0.714	-0.002	-1.440	-1.348	1.325	1.038
body_gyro_y	-0.001	-0.001	0.714	0.003	-1.441	-1.307	1.357	1.032
body_gyro_z	0.000	0.000	0.050	-0.019	0.014	-0.197	0.210	0.050

Table 6: Dataset Originale - Metriche riassuntive

Sensore	Media	Mediana	SD	Skewness	Kurtosis	Min	Max	MAD
body_acc_x	-0.001	-0.001	0.198	1.009	4.506	-1.232	1.300	0.024
body_acc_y	0.000	0.001	0.124	-0.990	5.938	-1.345	0.976	0.025
body_acc_z	0.000	0.000	0.109	-0.498	6.932	-1.365	1.067	0.026
body_gyro_x	0.001	0.000	0.416	-0.223	5.882	-4.734	4.155	0.086
body_gyro_y	0.000	-0.001	0.386	0.553	9.640	-5.974	5.746	0.077
body_gyro_z	0.000	0.001	0.260	-0.519	5.812	-2.763	2.366	0.063

## 7.5 Analisi Bivariata

Nell'analisi bivariata abbiamo calcolato la correlazione, tramite coefficiente di spearman, di ogni feature con l'attività, sono stati anche fatti degli scatterplot per vedere graficamente la relazione tra la feature ma hanno lo stesso andamento di quelli del dataset originale. Soffermiamoci a confrontare le correlazioni come sono variate nel dataset sintetico rispetto al dataset originale:

- Accelerazione asse X : nel dataset sintetico il coefficiente di correlazione è prossimo allo zero, mentre nel dataset originale la correlazione tra l'asse delle x è positivamente bassa.
- Accelerazione asse Y : l'accelerazione delle Y ha un coefficiente di correlazione pari a zero indicando che non è correlata con l'attività. Nel dataset originale l'accelerazione delle y è leggermente correlata negativamente con l'attività, indicando che man mano che l'attività aumenta, l'accelerazione diminuisce (le ultime attività comprendono quelle stazionarie, quindi man mano che l'attività sale l'accelerazione sulle y diminuisce).
- Accelazione asse Z : la correlazione dell'asse Z riporta lo stesso risultato in entrambi i dataset, inoltre essendo positiva vuol dire che all'aumentare dell'attività anche l'accelerazione sull'asse Z fa altrettanto.

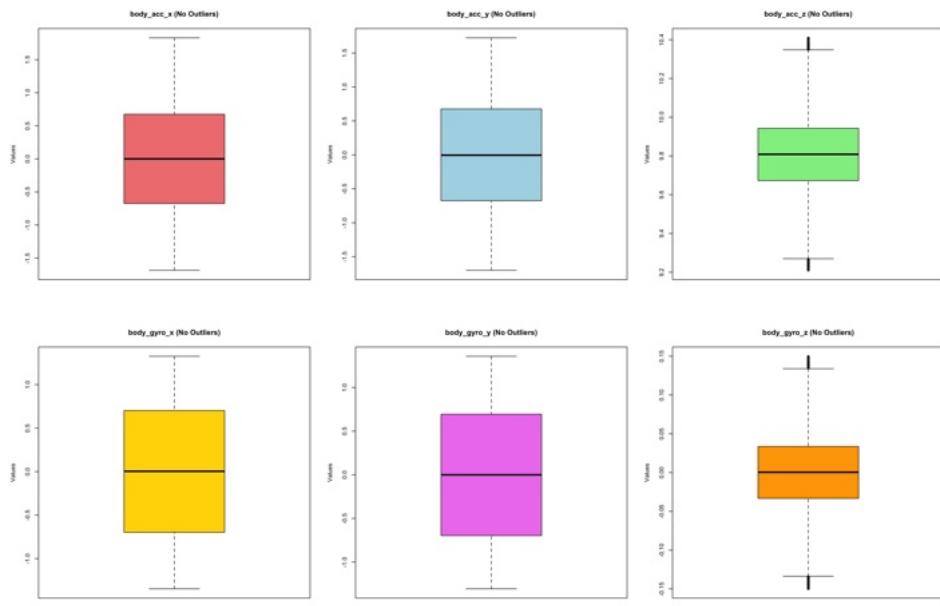


Figure 39: Dataset Sinetico - Boxplots delle variabili

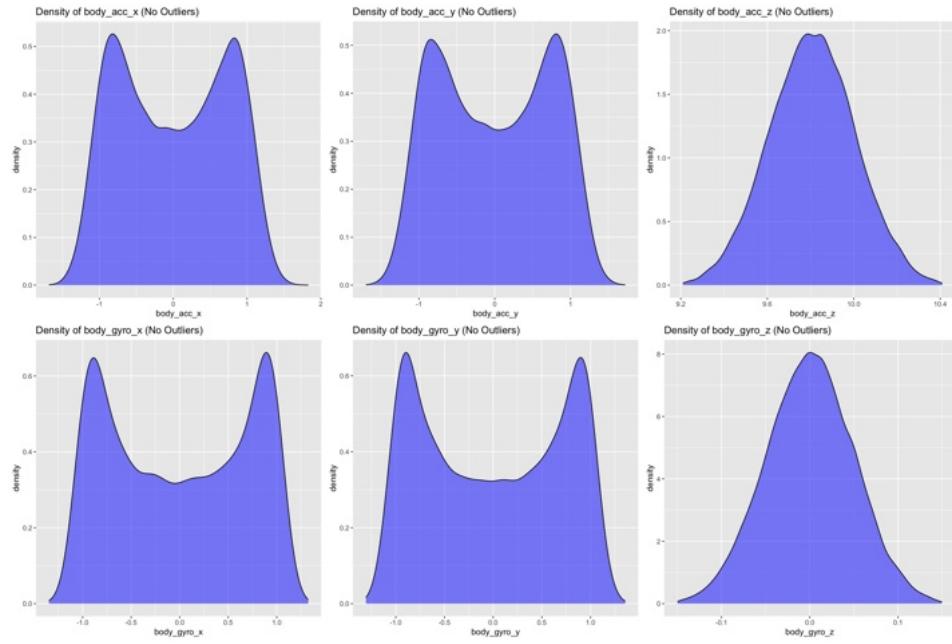


Figure 40: Dataset Sinetico - Distribuzione delle variabili

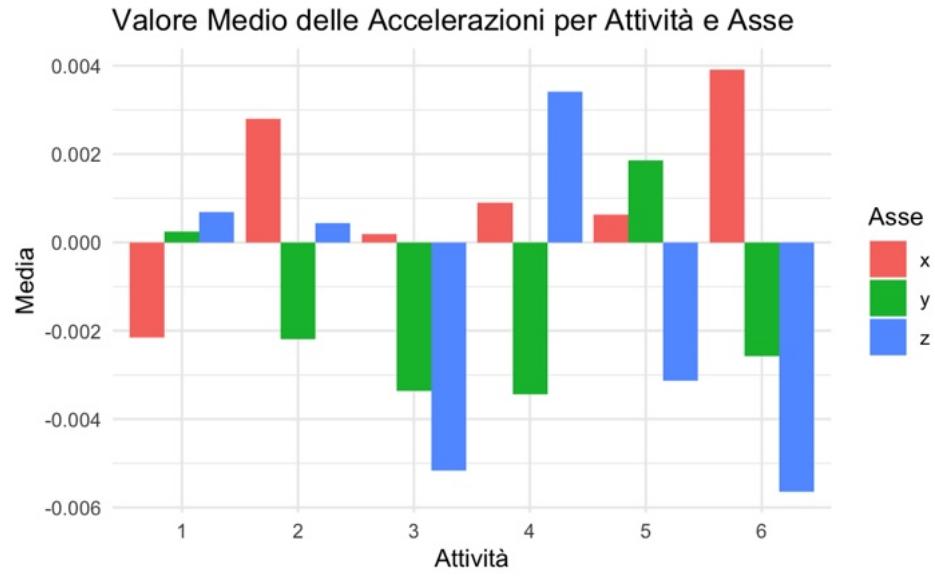


Figure 41: Dataset Sinetico - Valore medio delle accelerazioni per attività e asse.

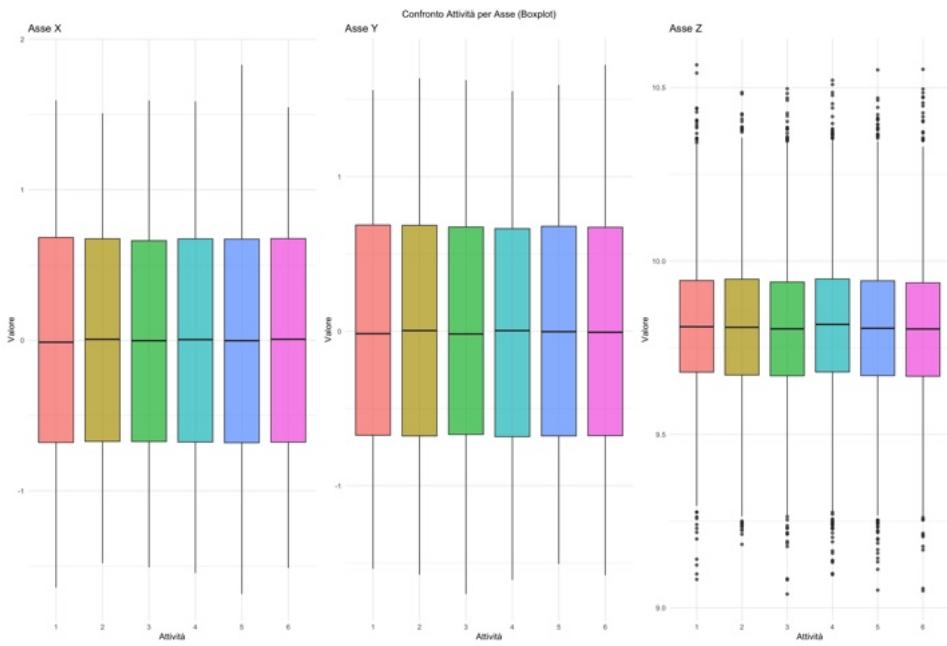


Figure 42: Dataset Sinetico - Boxplot di Attività per ogni Asse - Accelerazione

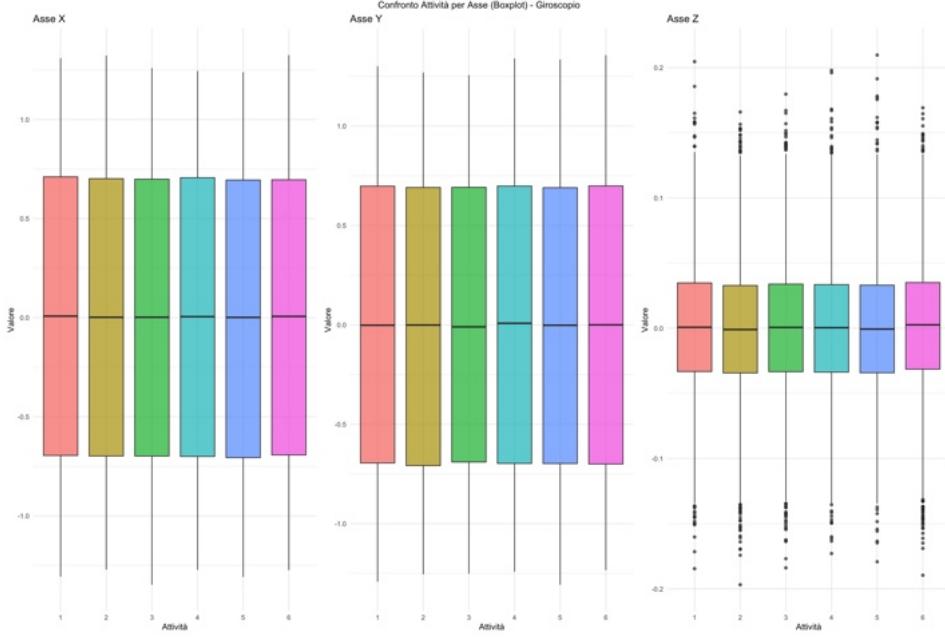


Figure 43: Dataset Sinetico - Boxplot di Attività per ogni Asse - Giroscopio

- Giroscopio asse X : per quanto riguarda la correlazione del giroscopio sull'asse X, nel dataset sintetico quest'ultima è leggermente, di pochissimo, correlata negativamente all'attività, mentre nel dataset originale abbiamo una correlazione leggermente, sempre di poco, positiva. In tal caso in entrambi i dataset il giroscopio tende a variare poco al variare dell'attività
- Giroscopio asse Y : il giroscopio dell'asse Y ha una leggera correlazione negativa con l'attività nel dataset sintetico, invece nel dataset originale, il giroscopio delle Y è correlato positivamente con l'attività anche se leggermente.
- Giroscopio asse Z : il giroscopio Z è leggermente correlato positivamente con l'attività mentre nel dataset sintetico è leggermente correlato negativamente.

Le altre variabili riguardanti l'accelerazione totale non sono presenti nel dataset sintetico generato da ChatGPT quindi non possiamo procedere ad un confronto con le correlazioni del dataset originale. Facendo riferimento alle correlazioni di queste feature nel dataset originale, sono quelle che presentano maggiore correlazione con l'attività, a differenza delle altre feature che nel dataset originale non hanno un altissima correlazione positiva o negativa con l'attività.

Variabile	Spearman
Body_Acc_X	0.002
Body_Acc_Y	0
Body_Acc_Z	1.00
Body_Gyro_X	-0.002
Body_Gyro_Y	-0.001
Body_Gyro_Z	0.006

Table 7: Coefficiente di correlazione tra le variabili e l'Attività

## 8 Sulla qualità degli LLM nel generare dataset numerici

In questa sezione rispondiamo indirettamente a due interrogativi fondamentali sull'uso dei Large Language Model (LLM) nella sintesi di dati sensoriali:

- **I LLM sono adatti per generare dati sintetici di accelerazioni lineari e velocità angolari?**
- **I dati sintetici da LLM possono sostituire i reali in task di clustering?**

### 8.1 Task di clustering

Per rispondere alle due domande, analizzeremo come si comporta il dataset sintetico nel rispondere alle prime due domande di ricerca

#### 8.1.1 È possibile determinare un'attività da una finestra di dati senza conoscere il paziente?

Per la verifica del primo obiettivo, abbiamo replicato la metodologia applicata al dataset originale. Inizialmente abbiamo tentato di distinguere le finestre di misurazione corrispondenti ad attività dinamiche (in movimento) da quelle statiche (a riposo). Tuttavia, come evidenziato nella Figura [44], non emerge alcuna separazione significativa tra le due categorie. Successivamente, abbiamo provato a classificare le singole attività all'interno dei gruppi dinamico e statico, ma i risultati sono stati insoddisfacenti (ARI 0.00 e V-Measure 0.002), valori compatibili con assegnazioni casuali.

Questo suggerisce che i dati generati da ChatGPT - modello linguistico addestrato su corpora testuali piuttosto che sulla sintesi di serie temporali numeriche - non preservino relazioni semantiche tra attività. Le finestre appaiono sostanzialmente casuali, prive dei pattern spaziotemporali necessari per discriminare anche macro-categorie come movimento/staticità.

#### 8.1.2 È possibile determinare il paziente da una finestra di dati conoscendo l'attività?

Analogamente alla verifica precedente, abbiamo adottato la stessa strategia del dataset originale: per ogni attività, abbiamo applicato KMeans impostando un numero di cluster pari ai pazienti disponibili [45]. Anche in questo caso, le metriche di valutazione (ARI 0.001 e V-Measure 0.001) indicano un'assenza di corrispondenza tra cluster generati e identità dei pazienti.

L'assenza di caratteristiche distintive riconducibili a specifici utenti nei dati sintetici rende impossibile qualsiasi forma di re-identificazione, confermando che le sequenze generate non catturano le peculiarità individuali presenti nei tracciati sensoriali reali. La totale assenza di pattern individuali (MAD inflazionata, distribuzioni omogenee) conferma che i LLM non catturano le "impronte digitali" cinematiche. I dati sintetici sono intercambiabili tra pazienti, invalidando qualsiasi studio biometrico.

Questi risultati forniscono una risposta operativa alle domande iniziali: i LLM attuali non sono strumenti adatti per la sintesi di dati cinetici, né possono sostituire dataset reali in task di clustering sensibile al contesto spaziotemporale.

## 8.2 Statistica inferenziale

**Le feature generate dai Large Language Model possono essere ricondotti ad una distribuzione normale?**

Durante quest'analisi, abbiamo verificato se una delle feature presenti è correlata con una distribuzione nota nella statistica. Per ogni variabile è stato applicato il test del chi-quadrato. La distribuzione che è stata considerata per ogni feature è stata la distribuzione normale o gaussiana. Per suddividere i valori di ogni feature in appropriati intervalli è stato utilizzato il metodo Freedam-Diaconis, dove IQR è l'intervallo interquartile ed N la dimensione del campione:

$$h = 2 \times \frac{IQR}{\sqrt[3]{N}}$$

Una volta calcolati gli intervalli appropriati, abbiamo calcolato la probabilità che la variabile potesse cadere in un intervallo, questo per ogni intervallo generato. Oltre alla probabilità, per il test chi-quadrato abbiamo calcolato le frequenze osservate e le frequenze attese che successivamente sono state per il test:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

**Dal testo di chi-quadrato è stato utilizzato un alpha pari a 0.05 e ha dato esito positivo solamente per l'accelerazione delle z e del giroscopio delle z, quindi solamente queste due feature si approssimano ad una normale.** Una volta verificato che l'accelerazione delle z e del giroscopio delle z, abbiamo calcolato gli intervalli di confidenza per i parametri di una normale utilizzando la media campionaria e la varianza campionaria che sono degli stimatori adatti. Siccome nella stima intervallare non si conoscevano né il mu né il sigma della variabile, si è utilizzato la varianza campionaria e fatto uso di una variabile di Studenti e una Chi-Quadrato.

$$VarX = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \approx T_{n-1}$$

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2}$$

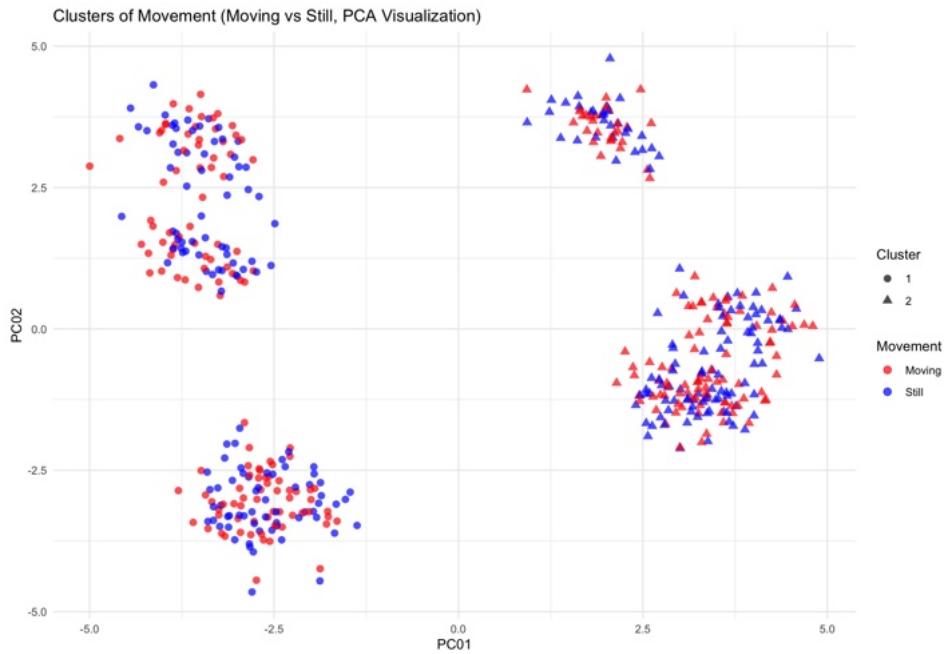


Figure 44: Dataset Sinetico - Moving vs Still

## References

- [1] R.-O. et Al., "Human Activity Recognition Using Smartphones," UCI Machine Learning Repository, 2013, DOI: <https://doi.org/10.24432/C54S4K>.



Figure 45: Clustering KMeans per la seconda ipotesi