# Conoscenza Strumenti Fintech

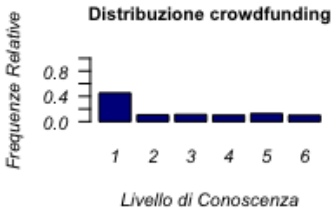
A. Cafiero, A. Cola, M. Simonetti, R. Urso

17 Aprile 2023

# Dataset e obiettivi

- Il dataset sul quale si intende condurre l'analisi riguarda un campione di 570 individui ai quali è stato sottoposto un questionario nell'ambito del progetto VMG "Differenze di genere nella conoscenza e nell'uso dei prodotti Fintech: esiste un ruolo per la trasparenza?"[1]
- In particolare, si è deciso di porre l'attenzione su 5 batterie di domande relative al livello percepito di conoscenza rispetto ad alcuni strumenti/oggetti finanziari. Per ogni item gli individui possono rispondere su una scala da 1 a 6, dove 1="non conoscenza" e 6="conoscenza perfetta".
- L'obiettivo dell'analisi è quello di utilizzare due algoritmi (EM ed iFCB) al fine di individuare gruppi omogenei di unità statistiche ed, eventualmente, caratterizzare tali gruppi mediante una serie di variabili socio-demografiche.

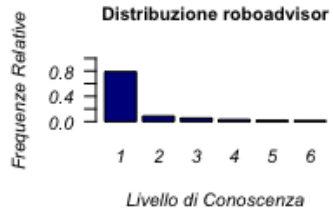
[1] Parte del progetto europeo CA19130 Fintech and Artificial Intelligence in Finance. Periodo di rilevazione: ottobre/novembre 2022.

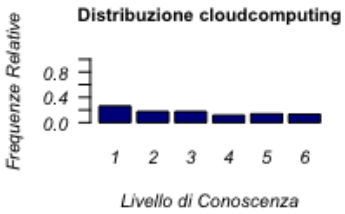

Distribuzione crowdfunding
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

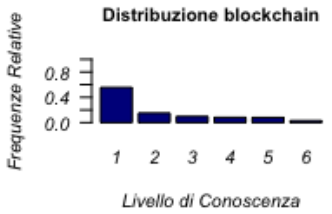

Distribuzione cryptocurrencies
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

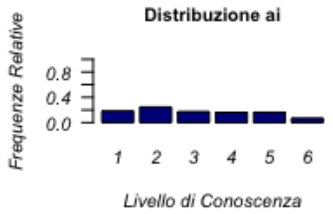

Distribuzione instantinsurance
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

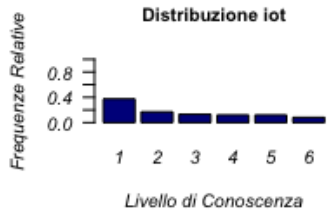

Distribuzione instantpayments
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione roboadvisor
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

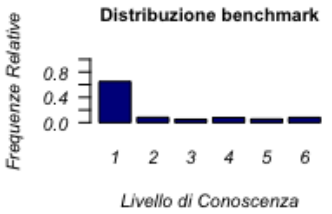

Distribuzione cloudcomputing
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione blockchain
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione ai
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

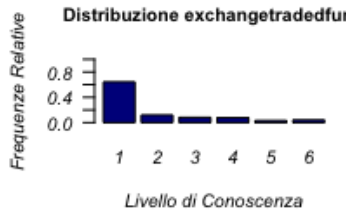

Distribuzione iot
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione machinelearning
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione benchmark
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione bonds
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione equity
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

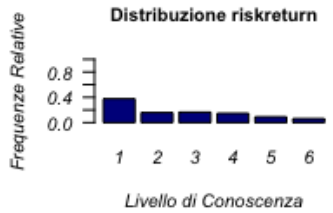

Distribuzione exchangetradedfund
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione investmentfund
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione crowdfundingopportunitie
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione financeopportunity
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

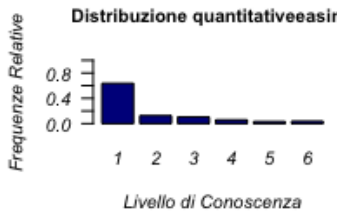

Distribuzione monetarypolicy
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

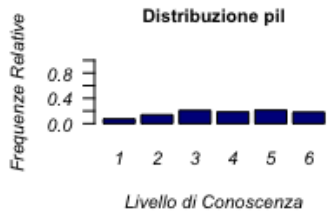

Distribuzione riskreturn
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

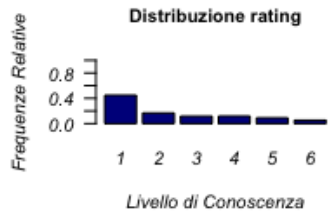

Distribuzione debitrisk
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione crisis
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione sustainablefinance
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione quantitativeeasing
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione pil
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza


Distribuzione rating
Frequenze Relative
0.8
0.4
0.0
1 2 3 4 5 6
Livello di Conoscenza

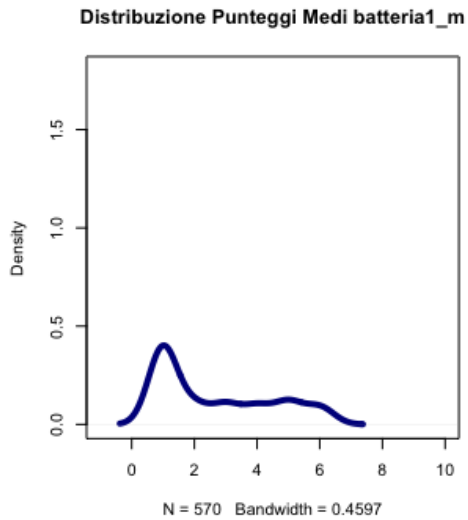
# $\alpha$ di Cronbach

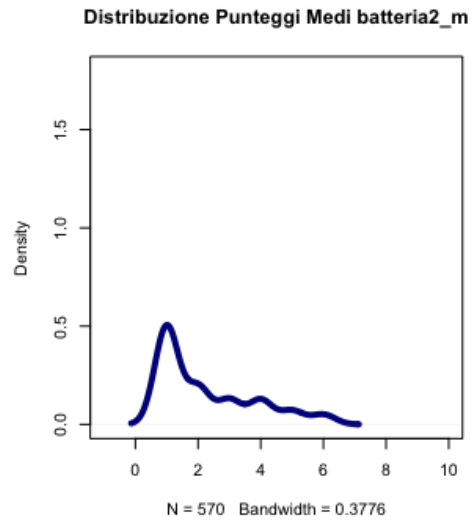
- Il primo passo è stato quello di utilizzare l'$\alpha$ di Cronbach per determinare il livello di coerenza degli item all'interno di ogni batteria.
- Per ognuna di queste è stato calcolato l'indicatore anche rimuovendo i singoli item, osservando come in nessuno di questi casi si rilevasse un incremento dell'$\alpha$.

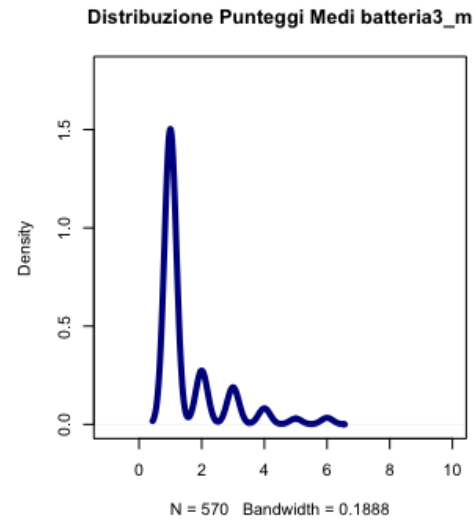
| Batteria1 | Batteria2 | Batteria3 | Batteria4 | Batteria5 |
|---|---|---|---|---|
| 0.82 | 0.88 | 0.94 | 0.93 | 0.93 |

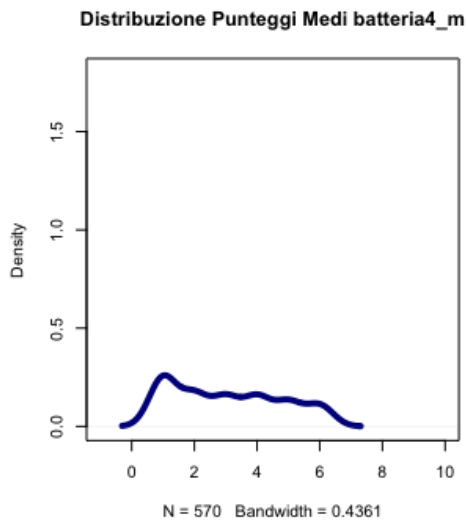
- Valutata la coerenza interna delle batterie, si può precedere estraendo da ognuna di essere una variabile ottenuta calcolando per ogni individuo il suo punteggio medio rispetto all'area di conoscenza rappresentata dalla specifica batteria.

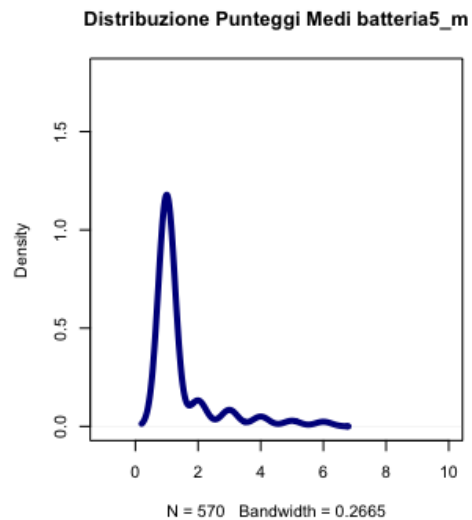
# Distribuzione punteggi medi



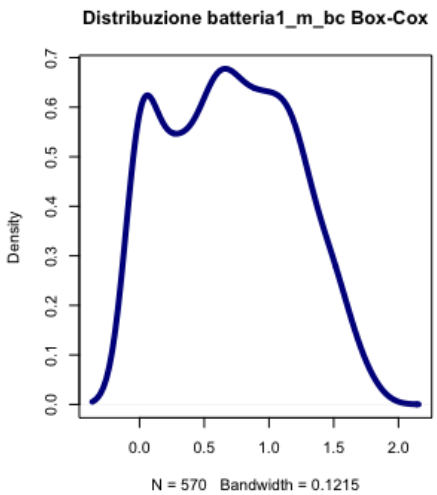



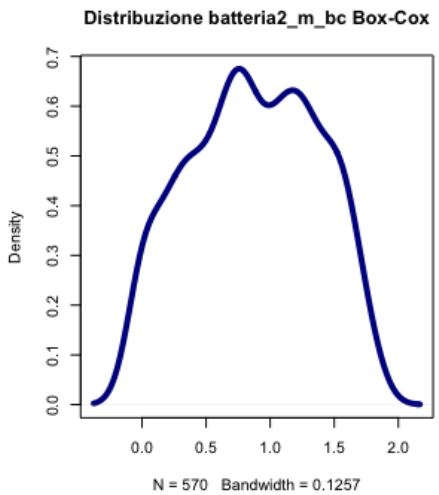



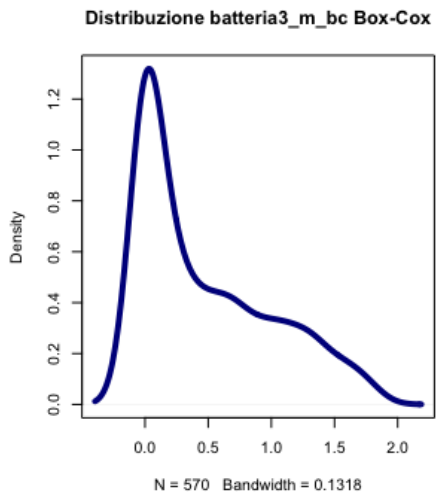



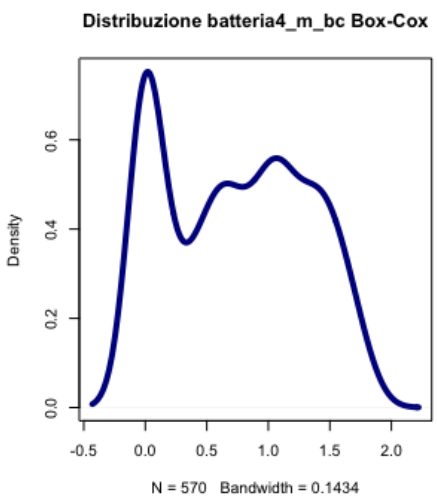
# Trasformazione Box-Cox

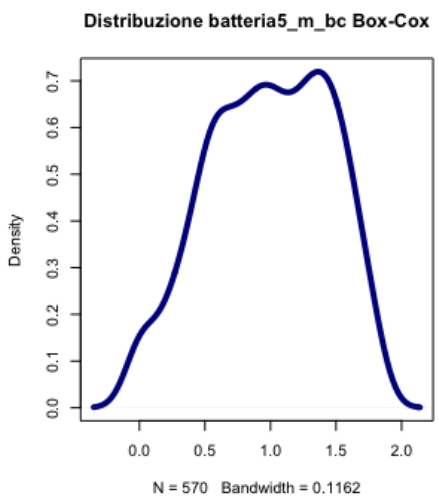
- Per normalizzare le distribuzioni dei punteggi è stata applicata una trasformazione logaritmica su tutte le variabili.



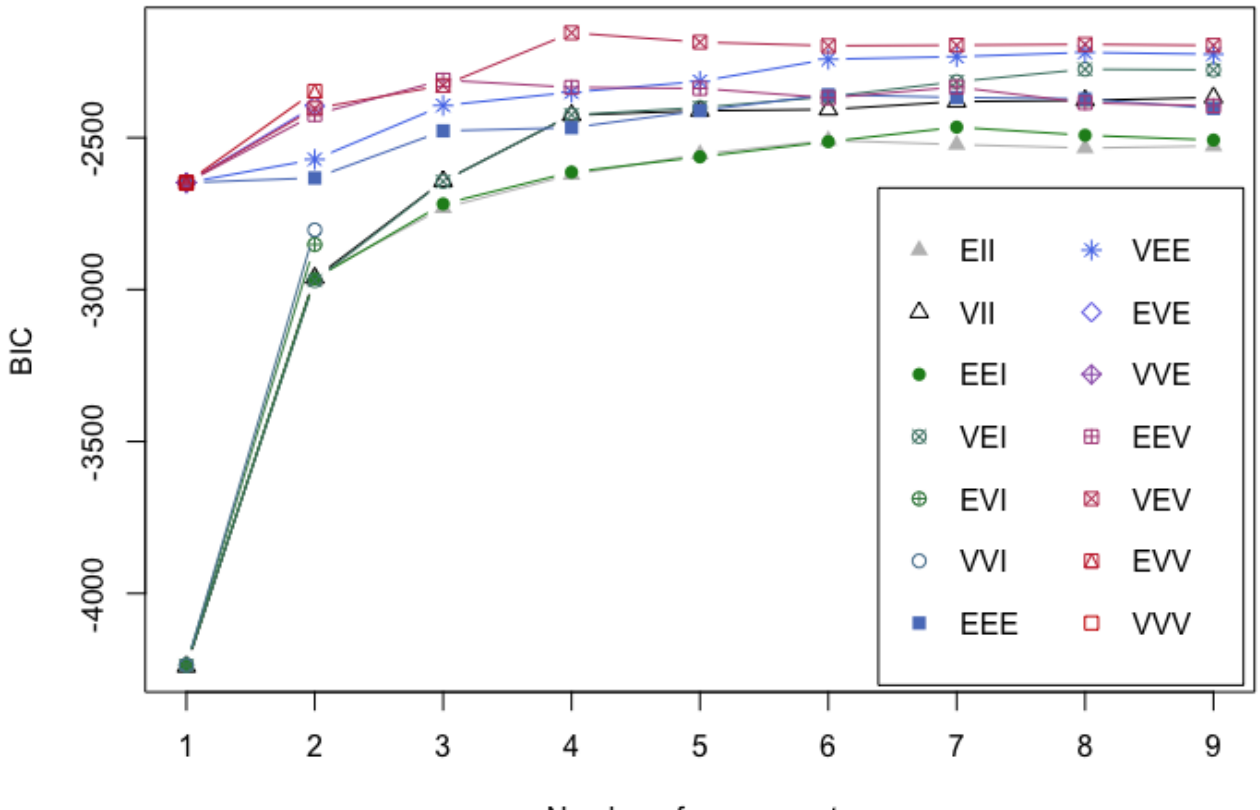



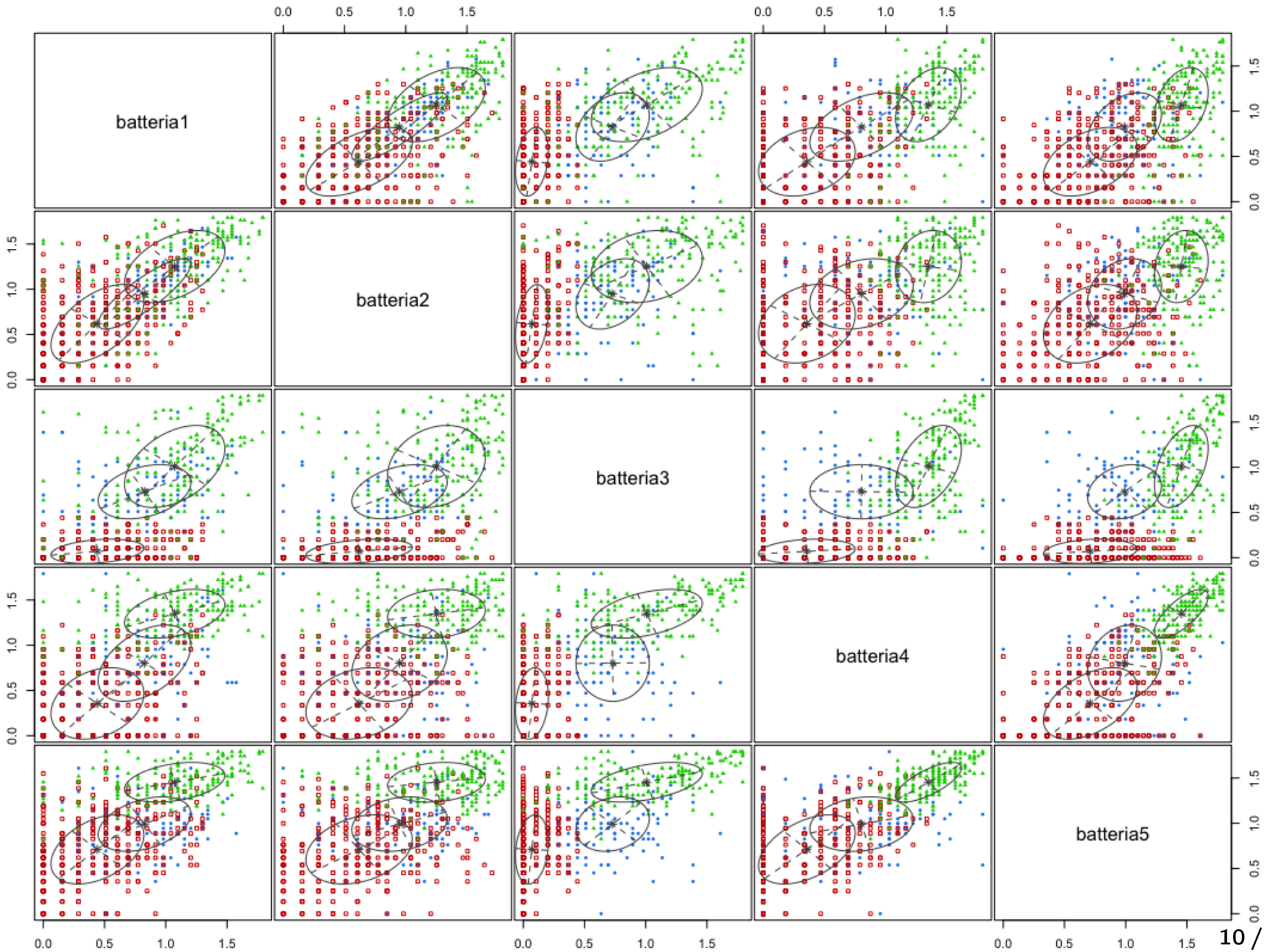



# Algoritmo EM

- Data la forma approssimativamente simmetrica delle distribuzioni, è stato applicato l'algoritmo EM al fine di stimare i parametri del miscuglio di gaussiane dal quale si suppone sia stato estratto il campione osservato dei punteggi medi.
- Il numero ottimale di cluster è stato individuato mediante un plot realizzato con il pacchetto *mclust*. Nel plot è riportato in ordinata il BIC di tutti i modelli stimati al variare del numero delle componenti del miscuglio (cluster).

- I modelli stimati si differenziano anche per i vincoli imposti sulla matrice di varianze e covarianze, ai quali sono riferite le sigle nel grafico.
- Da quest'ultimo si evince che il numero ottimale di cluster è pari a 4. Tuttavia, è stato ritenuto opportuno scegliere 3 componenti utilizzando un modello più parsimonioso.

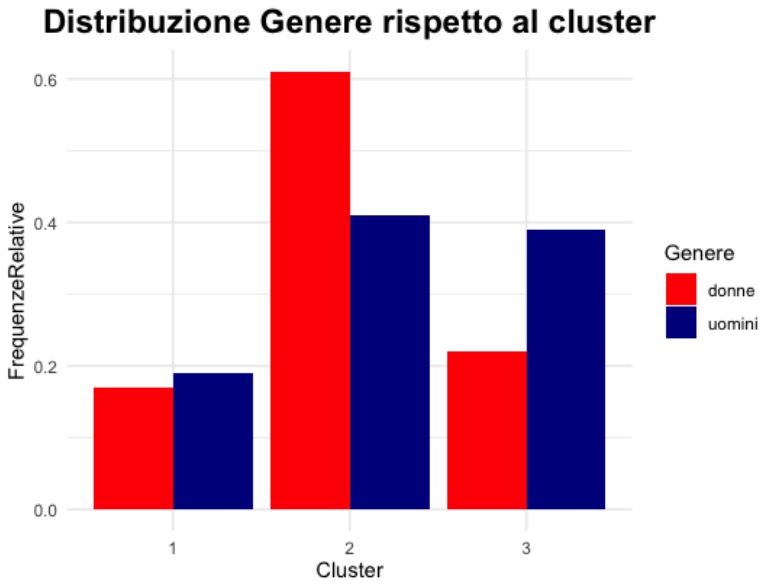


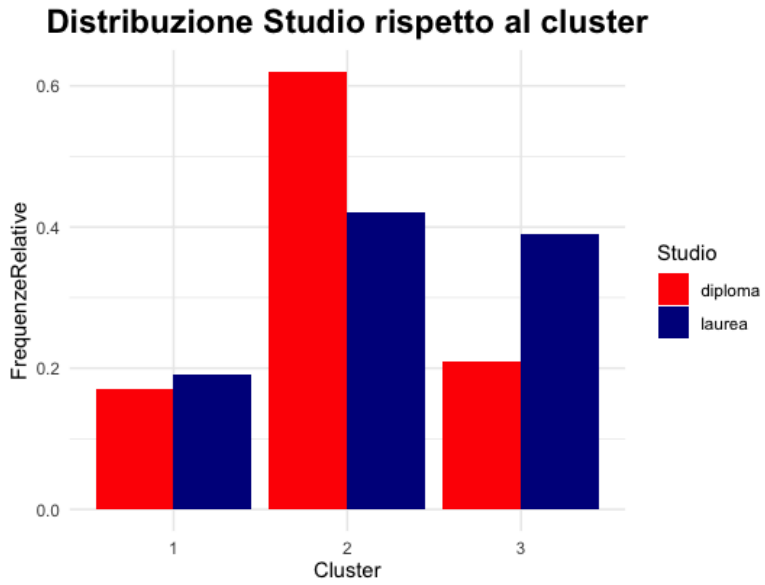

batteria1
batteria2
batteria3
batteria4
batteria5
0.0
0.5
1.0
1.5

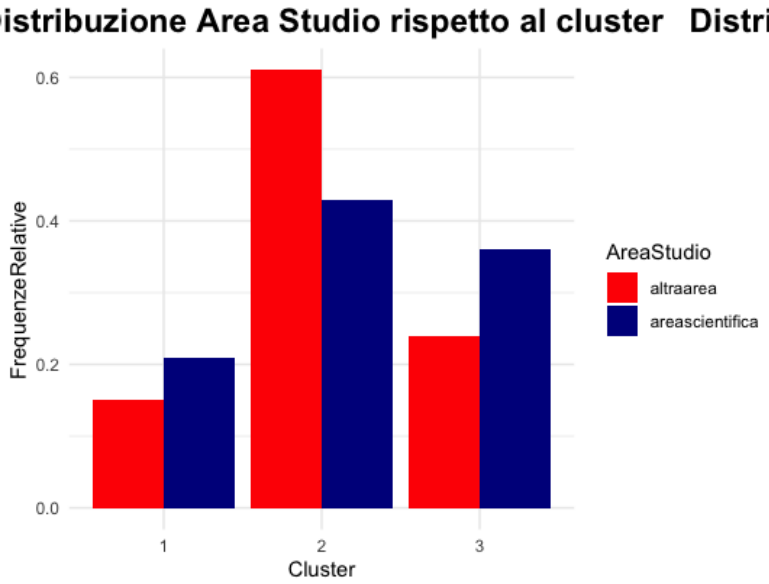
# Caratterizzazione cluster

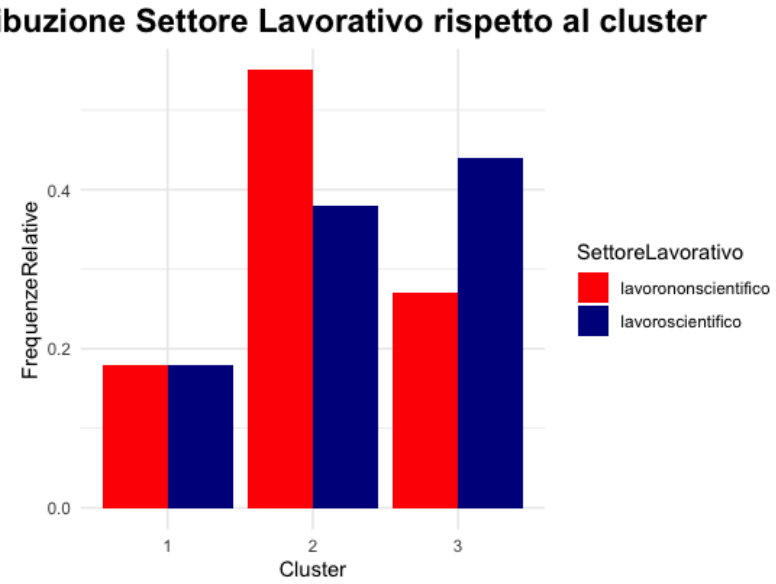
- Per caratterizzare i tre gruppi individuati è stato scelto un set di variabili socio-demografiche: genere, macroarea di residenza, titolo di studio, area disciplinare di quest'ultimo, lavoro, settore lavorativo.
- Per verificare se ci fosse o meno associazione tra la variabile di classificazione e le socio-demografiche, è stato effettuato un test $\chi^2$. Questo è risultato significativo solo per le variabili genere, titolo di studio, area disciplinare, settore lavorativo.
- Per ogni variabile è stato realizzato un grafico della distribuzione di frequenza rispetto ai tre cluster.

# Distribuzioni di frequenza rispetto ai gruppi



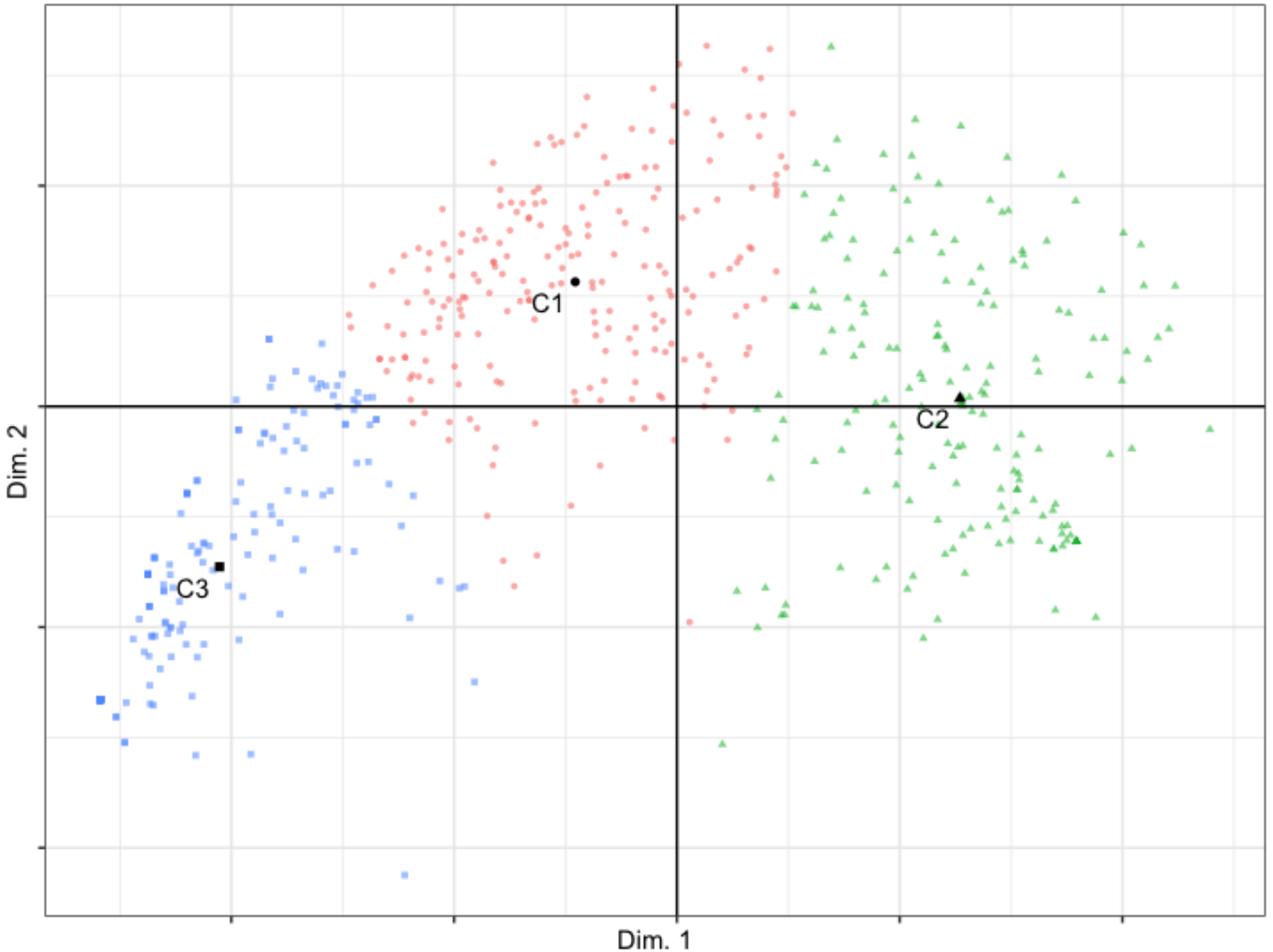
# iFCB

- È stato poi tentato un diverso approccio all'analisi, il quale prevede l'applicazione dell'algoritmo iFCB. Si tratta di una strategia integrata la quale opera una riduzione della dimensionalità e una classificazione nello spazio ridotto.
- Le variabili presenti nelle cinque batterie iniziali sono state dicotomizzate; in particolare sono state ricodificate ponendo:

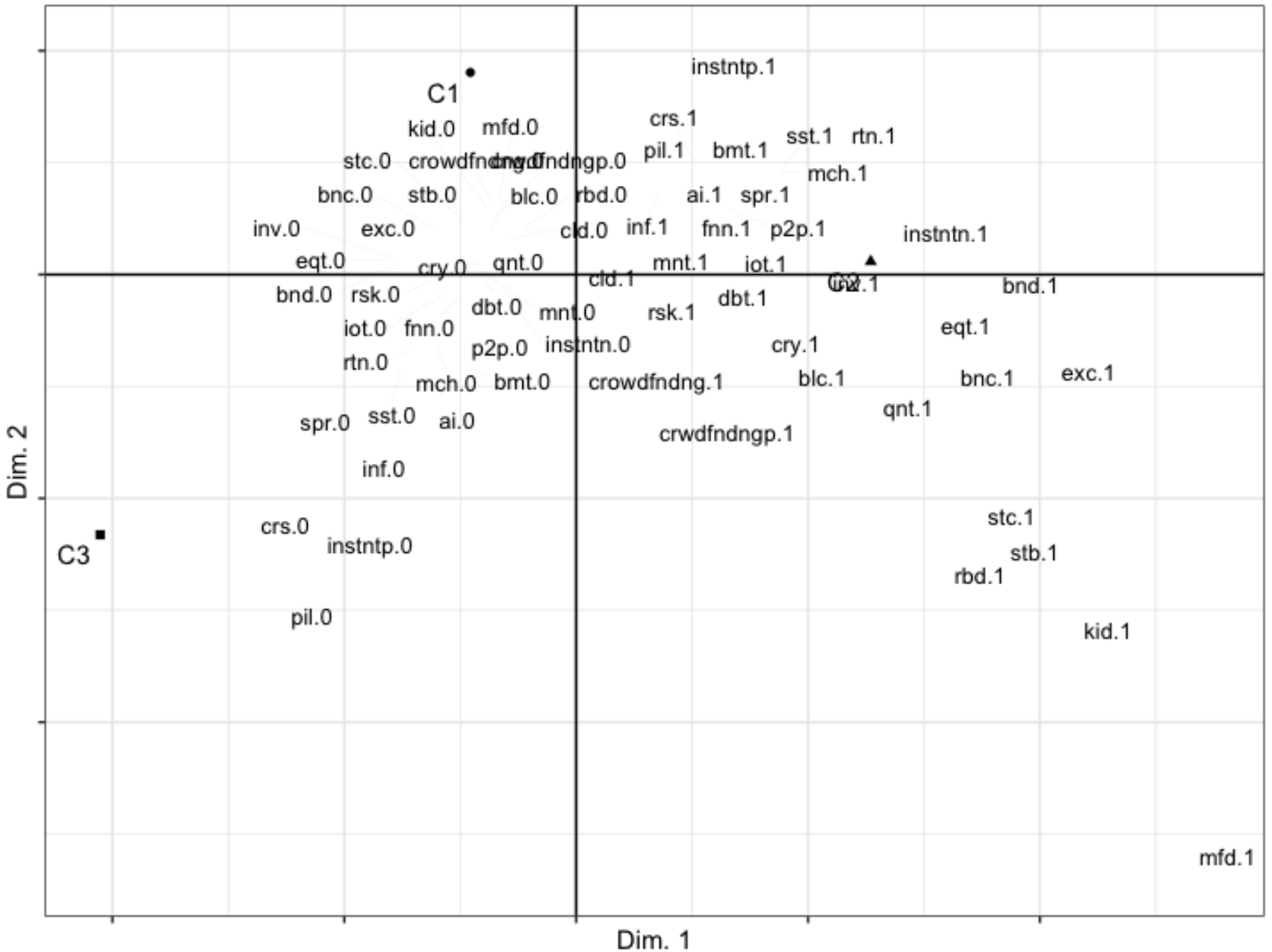
$$X = \begin{cases} 0 & \text{se il livello di conoscenza è pari ad 1} \\ 1 & \text{altrimenti.} \end{cases}$$

- Il pacchetto utilizzato per l'implementazione è *clustrd*. Il numero di cluster è stato fissato pari a tre.

# Unità nel piano fattoriale



# Modalità nel piano fattoriale



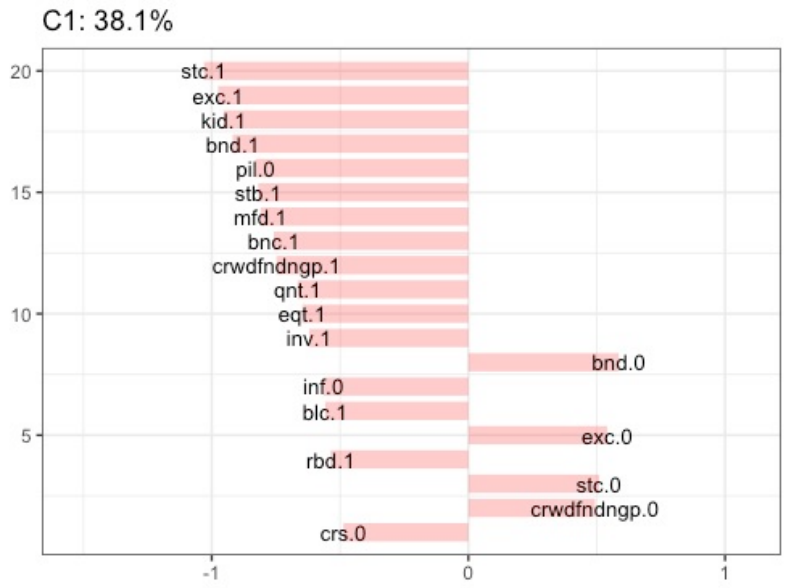
Ciò che emerge è la presenza di 3 gruppi così caratterizzati:

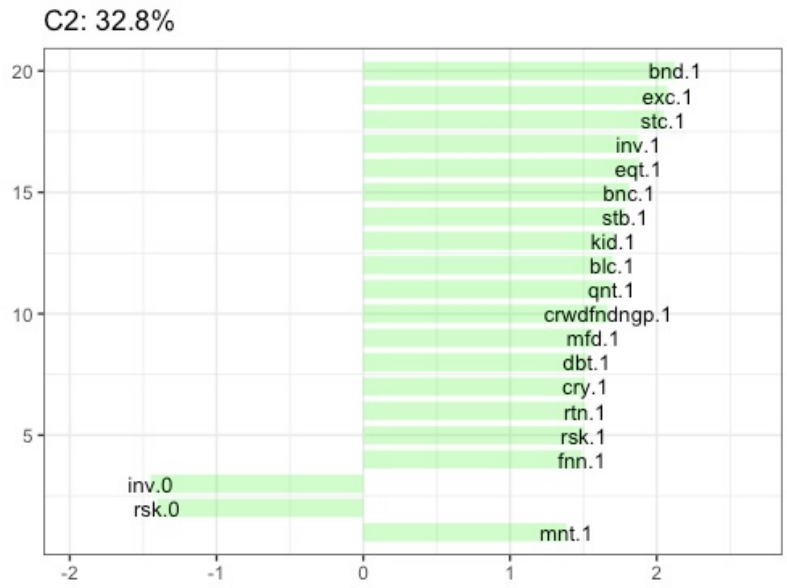
- Un gruppo di individui C3 che dichiara di non conoscere affatto argomenti piuttosto generici, tra cui PIL, inflazione, crisi;
- Il gruppo C2 afferma di conoscere sia strumenti/oggetti finanziari più specifici (MiFID, KID, Robo Advisor) che meno specifici (criptovalute, pagamenti istantanei, bond).
- Il gruppo C1 ignora argomenti più specifici ma dichiara di conoscere argomenti come l'intelligenza artificiale, lo spread, il cloud computing.

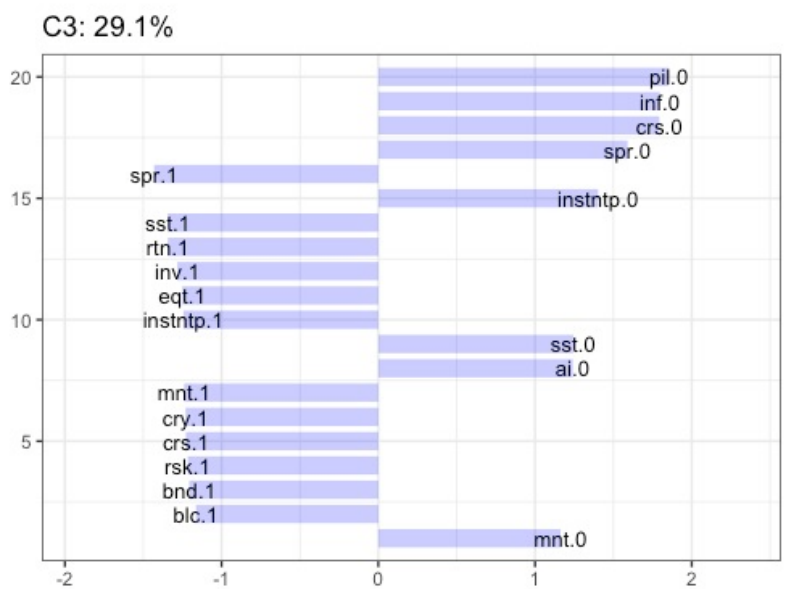
# Residui standardizzati

- Nella slide successiva sono riportati i barplot dei primi venti residui standardizzati, disposti per valore assoluto decrescente.
- Questi grafici sono utili per individuare quali sono le modalità che più si allontanano dalla condizione di indipendenza. Ai residui maggiori (in modulo) corrisponderanno le modalità che più caratterizzano lo specifico cluster.

# Residui standardizzati

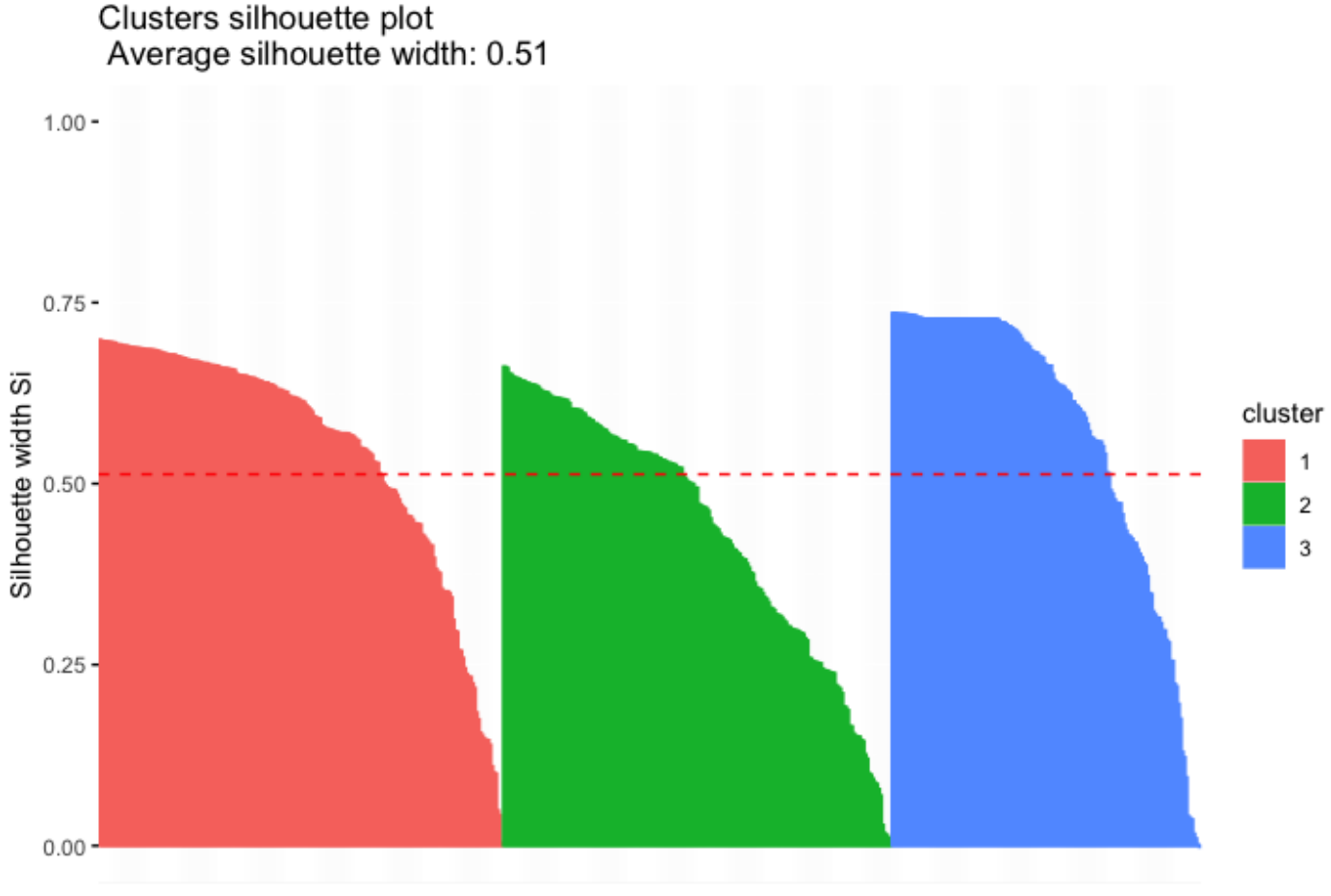


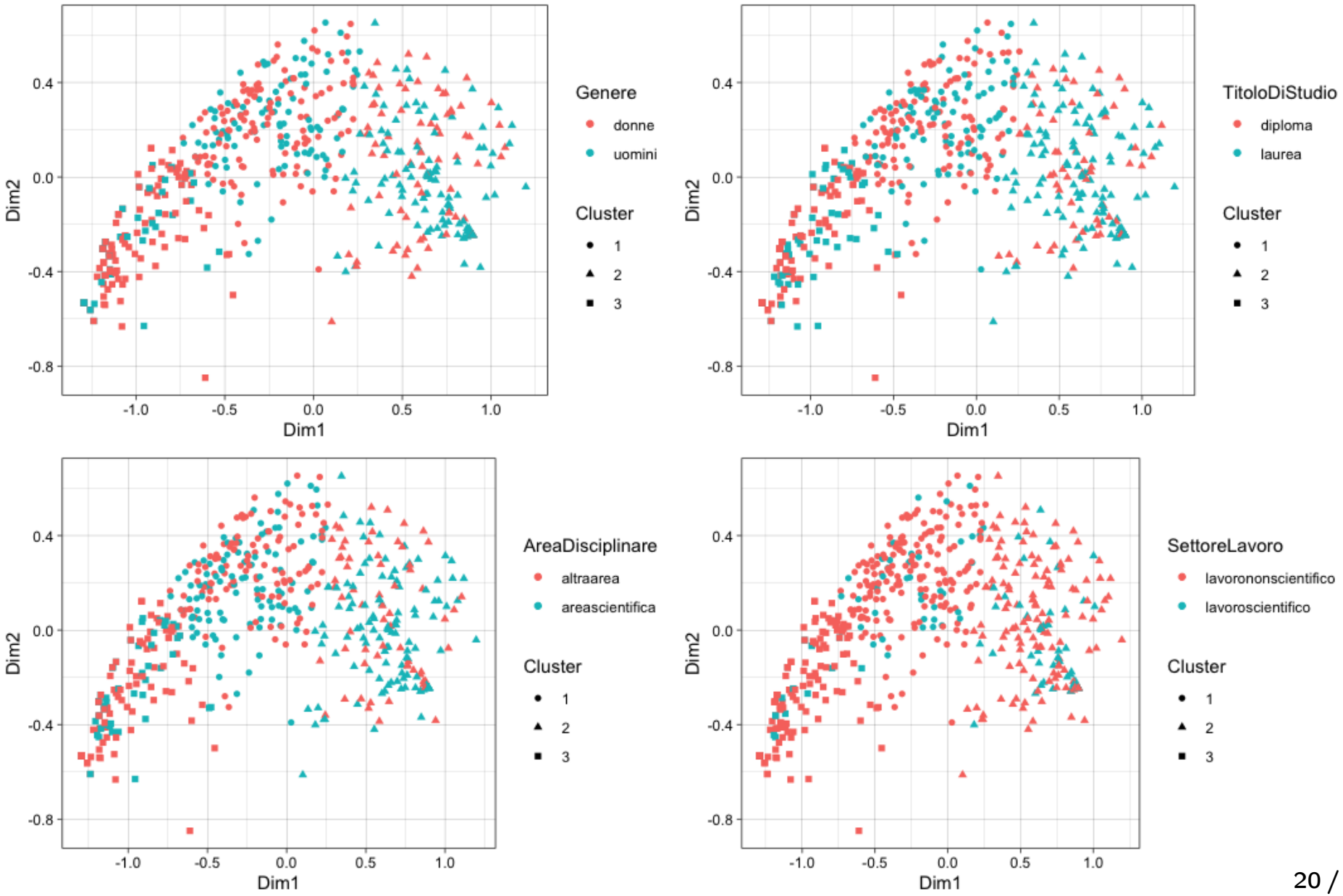



# Silhouette

- Per verificare se i gruppi sono stati ben separati è possibile utilizzare il seguente grafico. Per produrlo, viene calcolata la distanza di un punto da tutti gli altri. Il fatto che tutti i valori siano positivi significa che tutte le unità sono state ben classificate.



Anche per l'iFCB è stato effettuato un test $\chi^2$ per verificare l'associazione tra il set di variabili socio-demografiche e la variabile di classificazione.



# Conclusioni

- In conclusione, è stato osservato come entrambi gli algoritmi abbiano individuato un numero ottimale di gruppi pari a tre.
- I risultati del secondo approccio sono in linea con l'analisi precedente: dai dati emerge che ad essere in media più informati siano gli uomini, i laureati, coloro che hanno studiato in ambito scientifico e chi lavora in quest'ultimo settore.
- Rispetto alla caratterizzazione dei cluster, l'iFCB sembra fornire un'informazione aggiuntiva circa la specificità degli oggetti finanziari.

# Riferimenti


- Iodice D'Enza, A., Palumbo, F. (2013). Iterative factor clustering of binary data. Computational Statistics, 789-807.
- Markos, A., D'Enza, A. I., van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in R. Journal of Statistical Software, 91, 1-24.