Document: Applied srl

Creator: Ludovica d'Orsa

Supervisor: Ing. Alessandro Ruberti

Date: 14th January 2022

Connected cars: allegato tecnico

Obiettivi

Gli obiettivi del progetto sono:

- identificare delle tipologie di utenti utilizzando i dati raccolti dalle connected cars (Lamborghini Huracan),
- sviluppare una soluzione algoritmica che consenta l'associazione della vettura a una rispettiva tipologia utente.

Dati

I dati disponibili sono stati raccolti da 1457 connected cars, a partire dal 01 gennaio fino al 29 novembre 2021.

Dall'analisi sono state escluse 46 vetture di test e le vetture senza un minimo di tre giorni di record a velocità superiori a zero.

Metodologia

L'analisi si è svolta in cinque fasi distinte: pulizia dei dati, valorizzazione delle serie, selezione ed estrazione delle variabili, analisi dei cluster e validazione dei risultati.

1. Pulizia dei dati

Una volta ricevuti i dati con gli identificativi corretti, sono stati estratti e ristrutturati per garantire un formato che fosse adatto all'analisi.

2. Analisi serie

Essendoci spostamenti ritenuti poco utili ai fini della classificazione, si è rivelato necessario sviluppare un algoritmo che fosse in grado di distinguere le parti del dataset utili all'analisi dal resto del dataset.

L'algoritmo divide i dati in *serie*, periodo in cui una vettura passa da velocità zero a velocità superiori allo zero e torna ad essere ferma. La selezione delle serie si basa su due parametri:

- la velocità minima che deve raggiungere l'auto al suo interno perché la serie venga considerata valida,
- la distanza di tempo minima che deve intercorrere tra due serie perché vengano identificate come distinte.

Nel corso dell'analisi, i due parametri che sono stati scelti sono, rispettivamente,

- velocità media dell'automobile sommata a una deviazione standard dalla velocità,
- 5 minuti.

I parametri sono modificabili a seconda delle esigenze, del contesto e del tipo di analisi in corso.

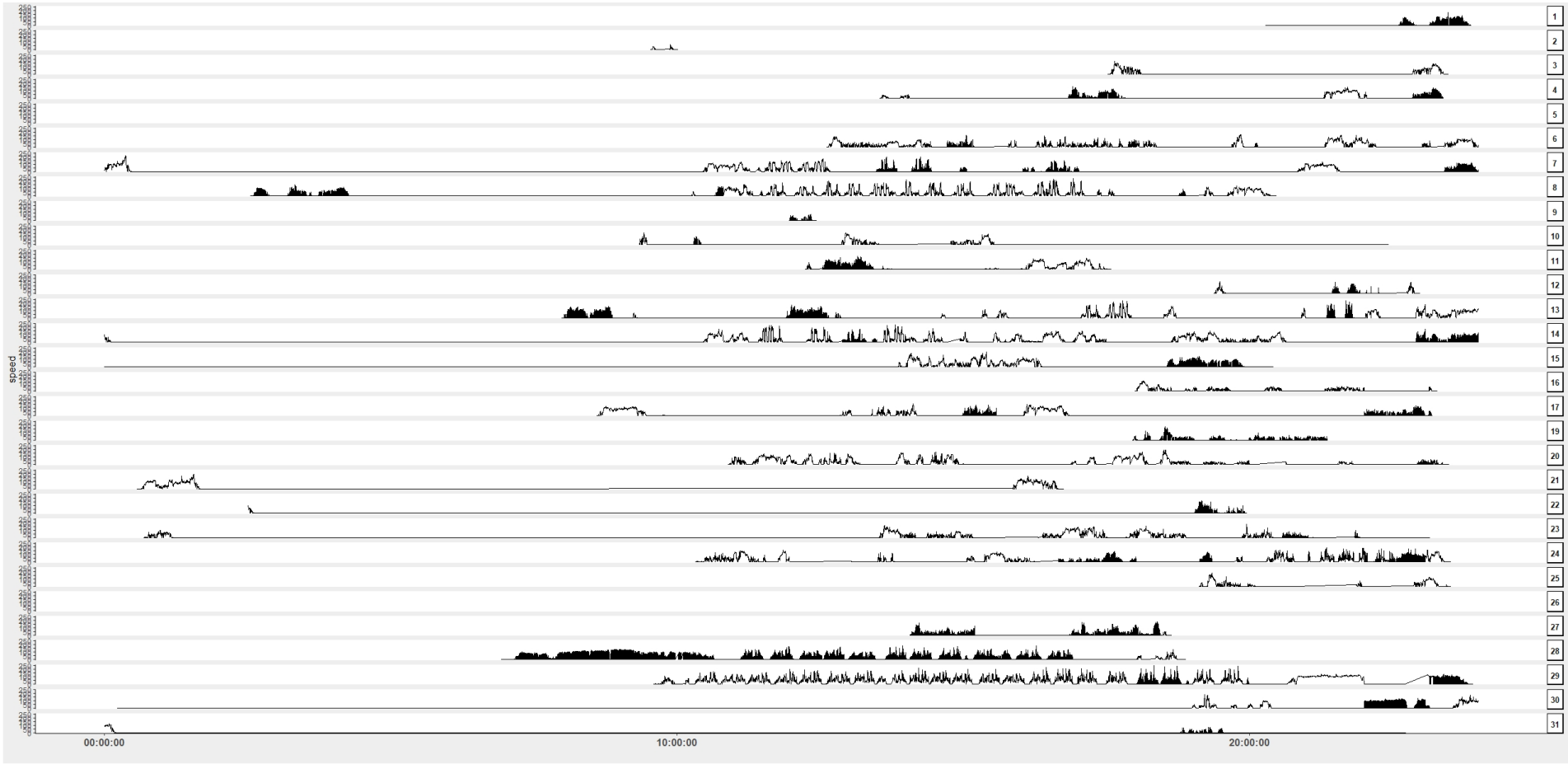

Figura 1: Rappresentazione degli spostamenti mensili di una vettura. Ciascuna riga orizzontale rappresenta un giorno del mese. Per ognuna di queste sono rappresentati l'orario sull'asse x e la velocità sull'asse y.

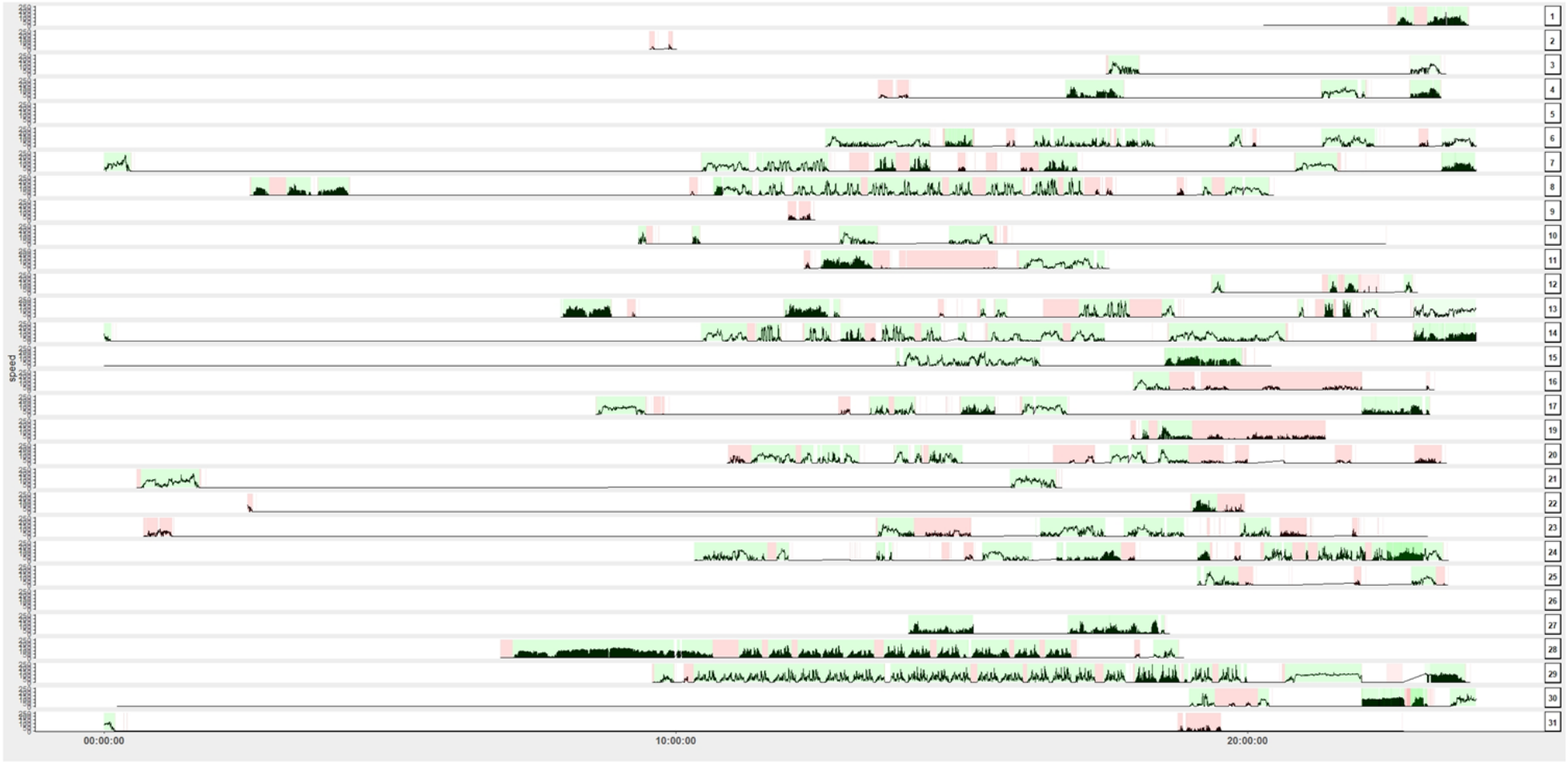

Figura 2: Rappresentazione degli spostamenti mensili di una vettura. Ciascuna riga orizzontale rappresenta un giorno del mese. Per ognuna di queste sono rappresentati l'orario sull'asse x e la velocità sull'asse y. Sono evidenziati in verde i periodi che l'algoritmo individua come utili all'analisi.

3. Analisi dei cluster

La metodologia scelta per identificare delle tipologie utenti è la cluster analysis, una classe di tecniche volte a identificare autonomamente raggruppamenti – cluster - in insiemi di dati sulla base di pattern. Permette di assegnare i veicoli a un gruppo di auto con caratteristiche simili senza dovere preimpostare delle categorie. Nello specifico, questa classe di algoritmi ha come obiettivo la minimizzazione della varianza totale intragruppo, concepita in termini di distanza in uno spazio multidimensionale. Per l'analisi è stato selezionato un algoritmo chiamato *k-means clusternig.* Per funzionare, prevede la selezione di un numero di cluster *k* e di delle variabili; queste scelte possono influenzare i risultati dell'analisi.

K-means opera selezionando casualmente k oggetti del dataset come punti centrali -*centroidi* - iniziali del cluster. Assegna ogni osservazione al centroide più vicino sulla base della distanza euclidea. Successivamente, per ognuno dei cluster cambia il centroide calcolando i nuovi valori medi di ogni oggetto nel cluster e ripete l'assegnazione dei centroidi e dei cluster fino a che l'assegnazione non smette di cambiare o viene raggiunto il numero massimo di iterazioni.

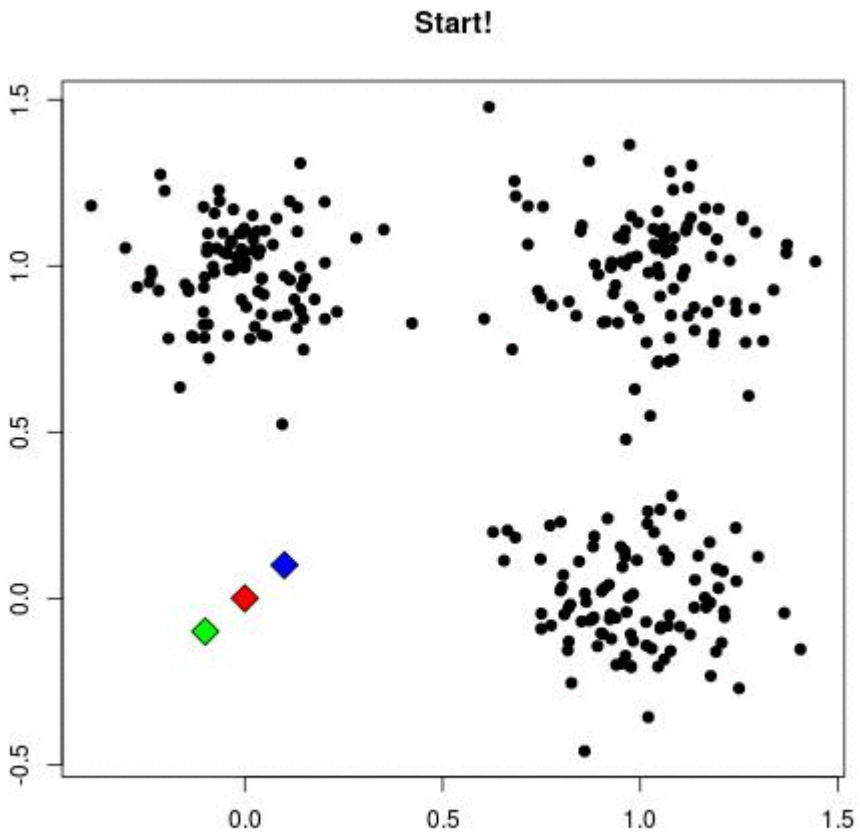

Figura 3: K-means clustering.

a. Scelta di k

La scelta del numero di *k* è stata guidata dai risultati ottenuti dall'applicazione ai dati di due tecniche empiriche: il metodo del gomito e gap statistics.

- Nel caso del metodo del gomito si itera l'algoritmo per diversi valori di *k* e si calcola la somma delle distanze tra ogni centroide ed i punti del relativo cluster al quadrato. La Figura 2 mostra sull'asse *x* i numeri dei potenziali cluster, sull'asse *y* la distanza calcolata. L'obiettivo degli algoritmi di clustering è minimizzare la varianza intragruppo, quindi la distanza tra i punti dello stesso cluser. Di conseguenza il cluster ottimale sarà quello dove la distanza tra centroide e punti del suo cluster è minima pur mantenendo un numero di *k* contenuto.

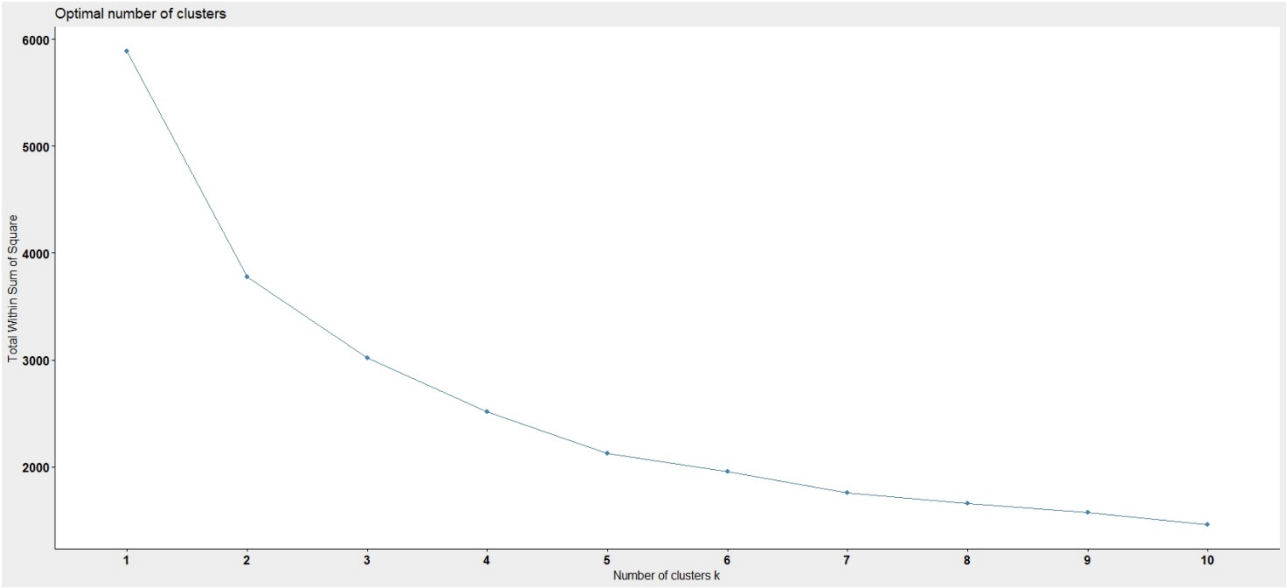

Figura 4: Metodo del gomito.

- In gap statistcs, la variazione totale interna al cluster per vari possibili valori di k viene comparata con i valori che ci si aspetterebbero in una distribuzione di dati senza cluster evidenti. Il cluster ottimale è quello in cui la distanza tra questi due valori, rappresentata sull'asse *y* del Grafico 3, è maggiore.

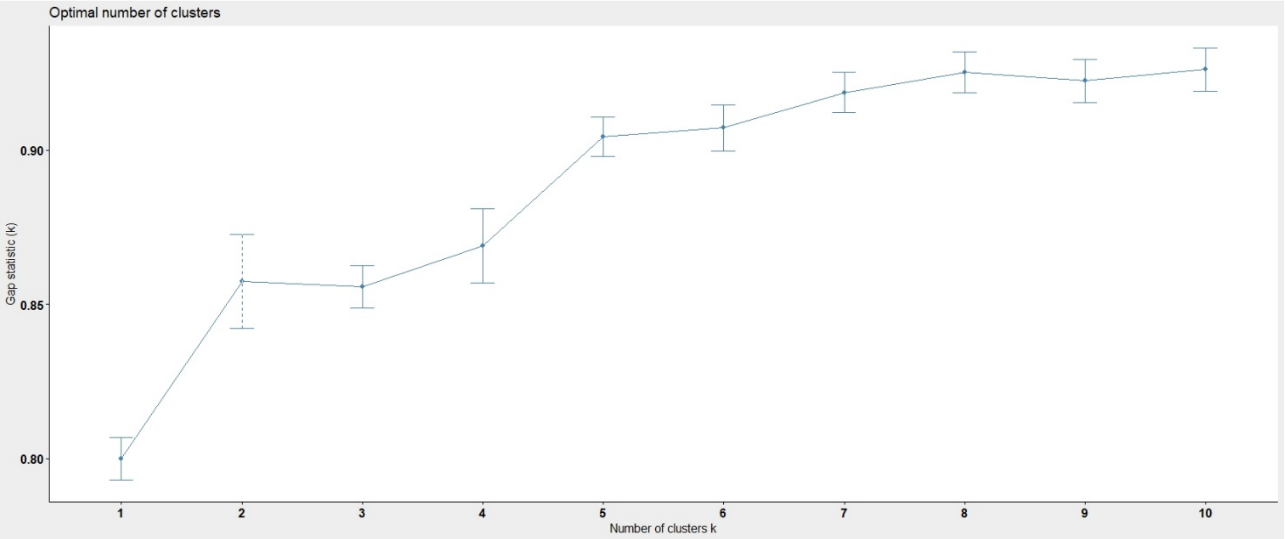

Figura 5: Gap statistics.

4. Scelta delle variabili

A causa della metodologia scelta per identificare delle tipologie utenti, è stato necessario selezionare delle variabili che fossero caratterizzanti e indicative delle abitudini degli utenti. Per questa analisi ne sono state selezionate cinque:

- **Frequenza di utilizzo dell'auto**, calcolata come frazione di giorni in cui si rilevano velocità superiori a zero rispetto al numero di giorni trascorsi tra la prima osservazione del dataset e il 29/11/21.
- **Velocità massima**, calcolata come media delle velocità massime quotidiane dell'auto.
- **Velocità media.**
- **Chilometri totali percorsi dalla vettura**, calcolati come differenza tra il valore segnato dal contachilometri nell'ultima osservazione e quello della prima osservazione.
- **Durata media di uso della vettura**. Una volta individuate le serie valide, la soluzione algoritmica calcola la durata delle serie, ne somma la durata su base giornaliera e ne calcola la media.

5. Findings

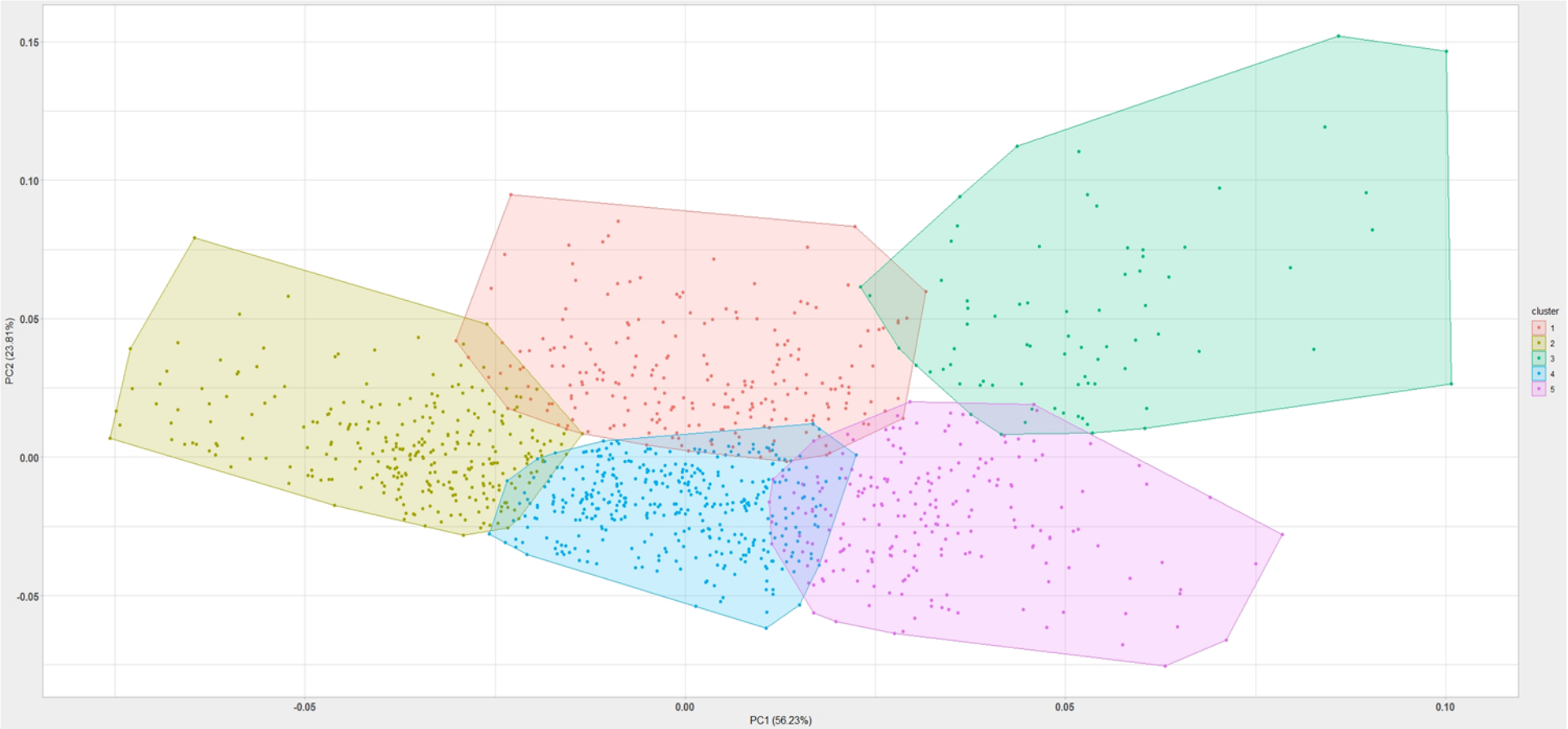

Figura 6: Risultati della cluster analysis.

6. Validazione dei risultati

L’analisi dei cluster si basa su un’attività di apprendimento automatico non supervisionato: non ci sono categorie preimpostate con le quali confrontare i risultati del clustering effettuato; quindi, non è possibile verificare matematicamente l’accuratezza dei risultati. Di conseguenza, è stato necessario adottare una metodologia alternativa per la validazione, basata sul confronto tra i gruppi ottenuti e variabili del dataset che non sono state utilizzate per la clusterizzazione.

In questo modo è possibile verificare se i cluster mostrano tratti distintivi per quanto riguarda una gamma più ampia di comportamenti.