Connected Cars

23/12/2021

Agenda

- Obiettivi e dati
- Metodologia
- Valorizzazione serie
- Cluster analysis
- Selezione variabili
- Findings
- Validazione dei risultati
- Possibili applicazioni

Contenuti

Dati:

- Dal 01 gennaio al 29 novembre 2021
- 1457 vetture, di cui 46 di test
- 481,762,279 osservazioni totali
- 882 auto in Europa, 438 in America, 113 in Asia, 9 nel resto del mondo e 15 la cui posizione è sconosciuta

Obiettivi:

- Identificare delle tipologie di utenti utilizzando i dati raccolti dalle connected cars
- Sviluppare una soluzione algoritmica che consenta l'associazione della vettura a una tipologia utente

Metodologia

01

Pulizia dei dati
Ristrutturazione ed estrazione dei dati

02

Valorizzazione serie
Divisione del dataset in serie e individuazione delle serie utili per l'analisi

03

Selezione variabili
Identificazione delle variabili che meglio caratterizzano insiemi di utenti

04

Cluster analysis
Categorizzazione delle vetture tramite tecniche di clustering

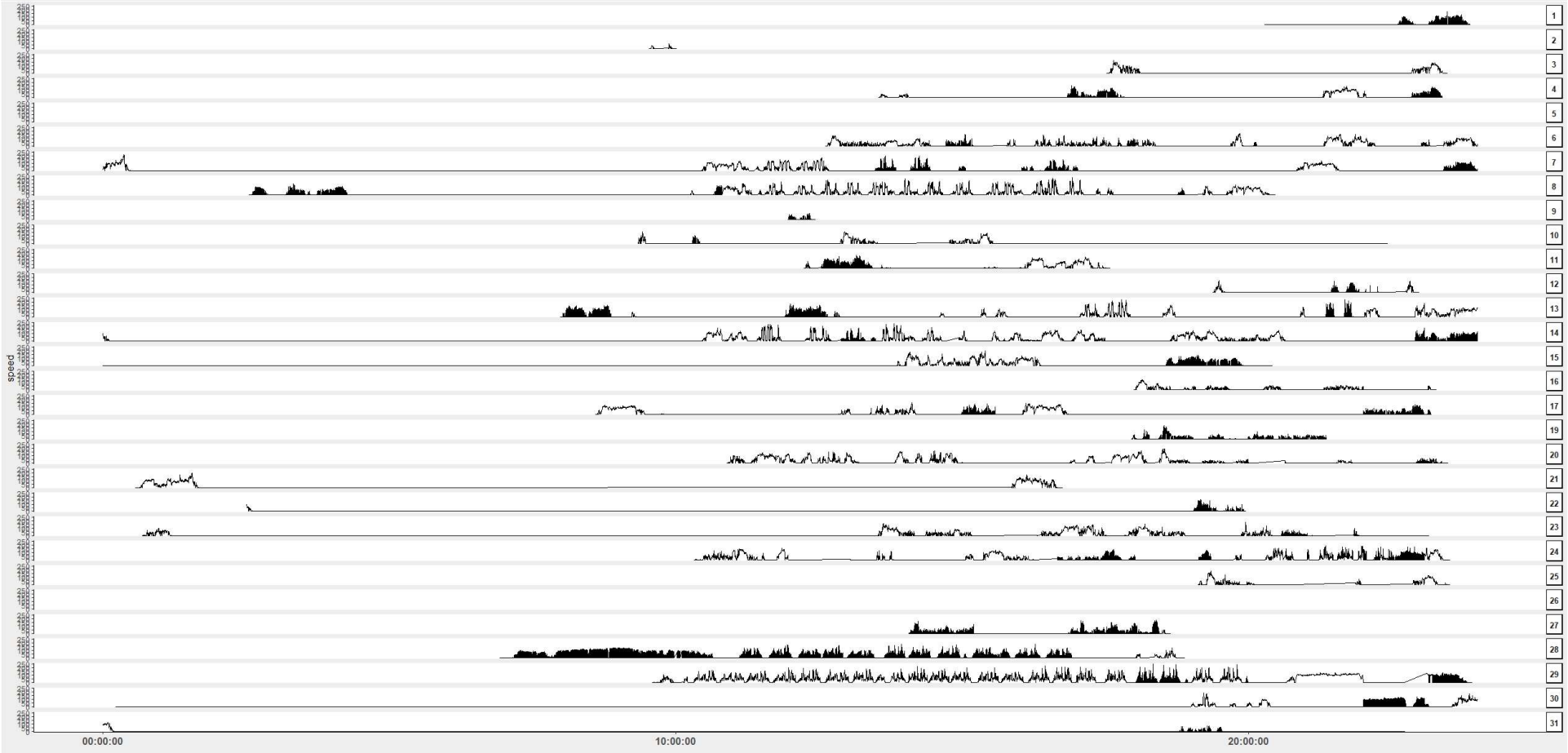
05

Findings
Analisi, interpretazione e validazione dei risultati

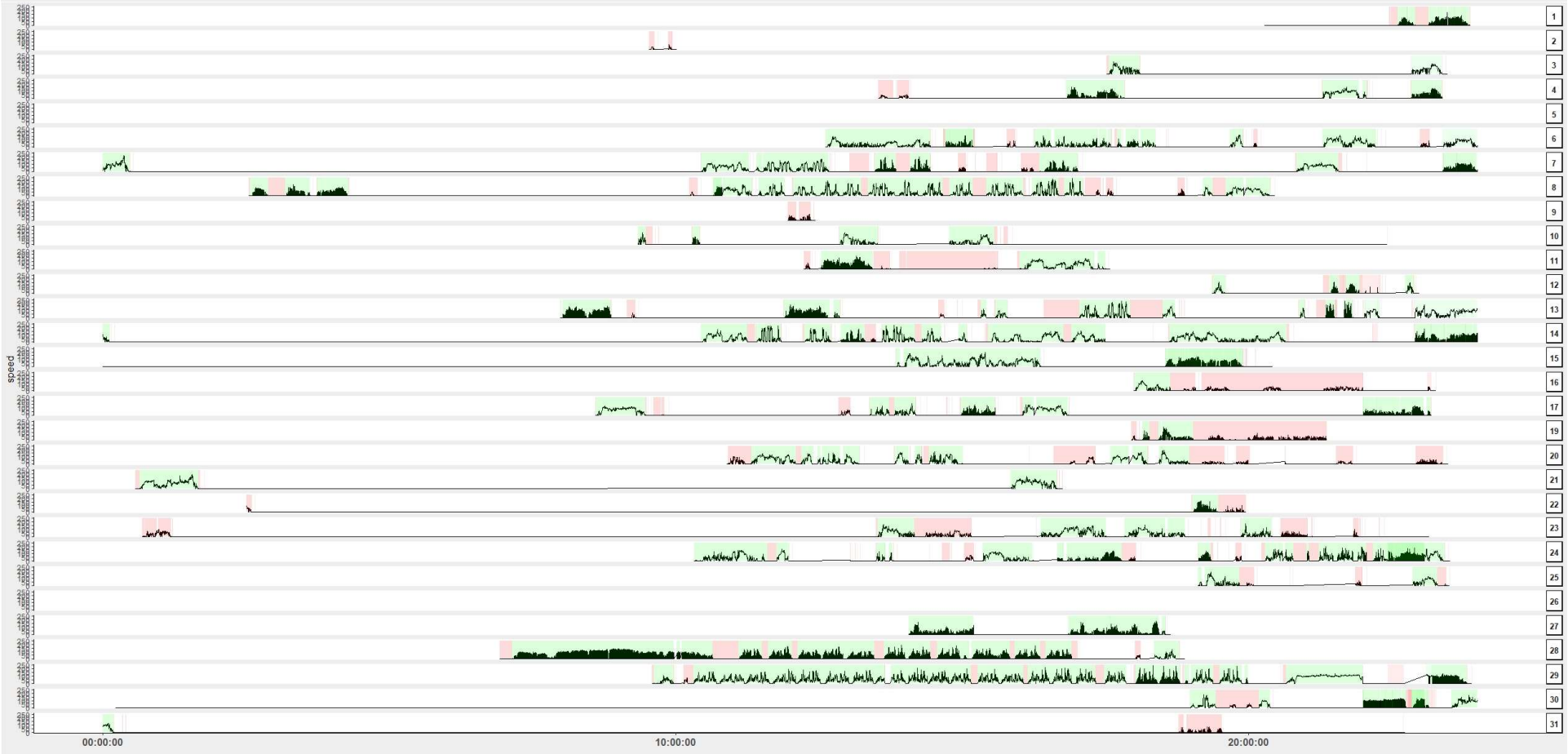
Valorizzazione serie

- Per riuscire a trarre informazioni significative dai dati è necessario distinguere i periodi utili all'analisi dal resto del dataset
- Per farlo è stato sviluppato un apposito **algoritmo**
- L'algoritmo divide il dataset in **serie:** periodi in cui la macchina passa da una velocità uguale a zero a velocità superiori e torna a zero.
- La selezione delle serie è basata su due **parametri**:
    - La **velocità** che deve raggiungere al suo interno la vettura perché la serie sia considerata valida
    - La **distanza di tempo** che deve intercorrere tra due serie perché siano considerate distinte

Valorizzazione serie



Valorizzazione serie



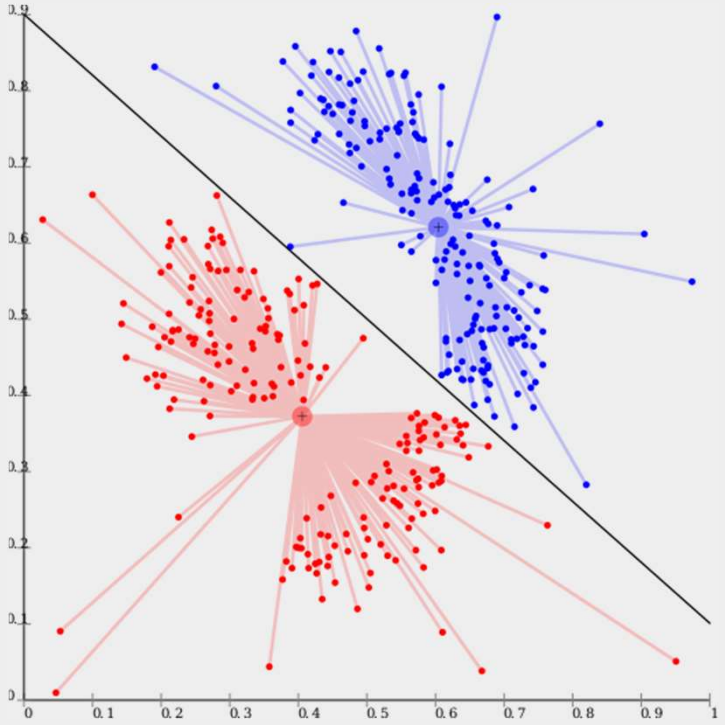
Cluster analysis

- L’analisi di cluster prevede la scoperta autonoma di raggruppamenti (**cluster**) nei dati sulla base di pattern
- Nello specifico, questa classe di algoritmi ha come obiettivo la minimizzazione della varianza totale intra-gruppo, concepita in termini di distanza in uno spazio multidimensionale
- Questo tipo di tecnica statistica permette di assegnare i veicoli a un gruppo di auto simili sulla base di caratteristiche comuni senza bisogno di preimpostare delle categorie
- Si tratta di un’attività di apprendimento automatico **non supervisionato**
- Al contrario di algoritmi di apprendimento supervisionato, i risultati non possono essere confrontati con delle categorie preassegnate

Cluster analysis

K-means clustering

- **K-means** è l'algoritmo di clusterizzazione selezionato per l'analisi
- Prevede che si selezioni un numero di cluster (**k**) prima dell'analisi, questa scelta influenza i risultati
- Abbiamo ipotizzato l'esistenza di **5 cluster** e confermato l'ipotesi tramite metodi empirici (in annex)
- Come ogni algoritmo, richiede delle variabili come input

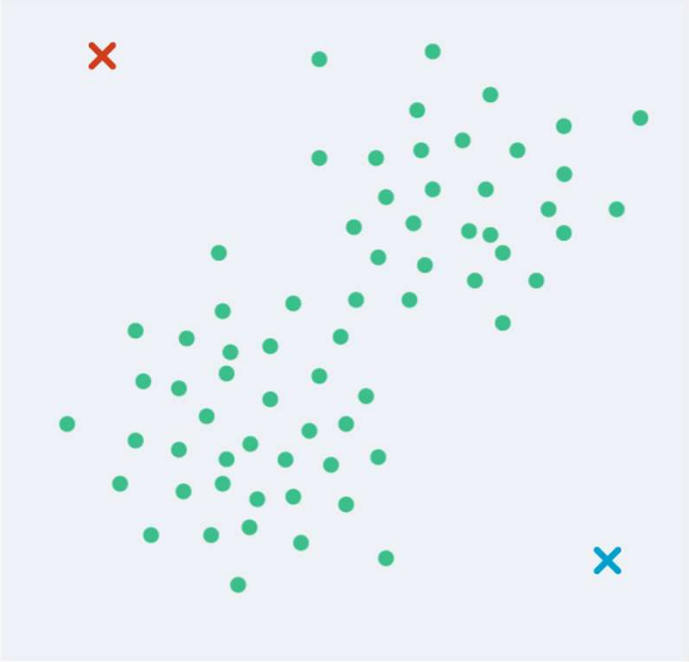


Cluster analysis

K-means clustering

Funzionamento:

1. Seleziona casualmente k oggetti del dataset come centri (**centroidi**) iniziali del cluster
2. Assegna ogni osservazione al centroide più vicino sulla base della distanza euclidea
3. Per ognuno dei cluster cambia il centroide calcolando i nuovi valori medi di ogni oggetto nel cluster.
4. Ripete i punti 3 e 4 fino a che l'assegnazione dei punti non smette di cambiare o viene raggiunto il numero massimo di iterazioni.



Cluster analysis

Variabili

La divisione in cluster non è stata eseguita sulla base dell'intero dataset, bensì su un numero di variabili selezionate in quanto più significative e adatte a descrivere le abitudini degli utenti:

1. **Frequenza di utilizzo dell'auto**

Calcolata come frazione di giorni in cui si rilevano velocità superiori a zero rispetto al numero di giorni tra la prima osservazione del dataset e il 29/11/21.

2. **Velocità massima raggiunta**

Calcolata come la media delle velocità massime quotidiane dell'auto.

3. **Velocità media**
4. **Chilometri totali percorsi**

Calcolata come differenza tra il valore segnato dal contachilometri nell'ultima osservazione e quello della prima osservazione.

5. **Durata media di uso della vettura**

Cluster analysis

Variabili

5. Durata media di uso della vettura

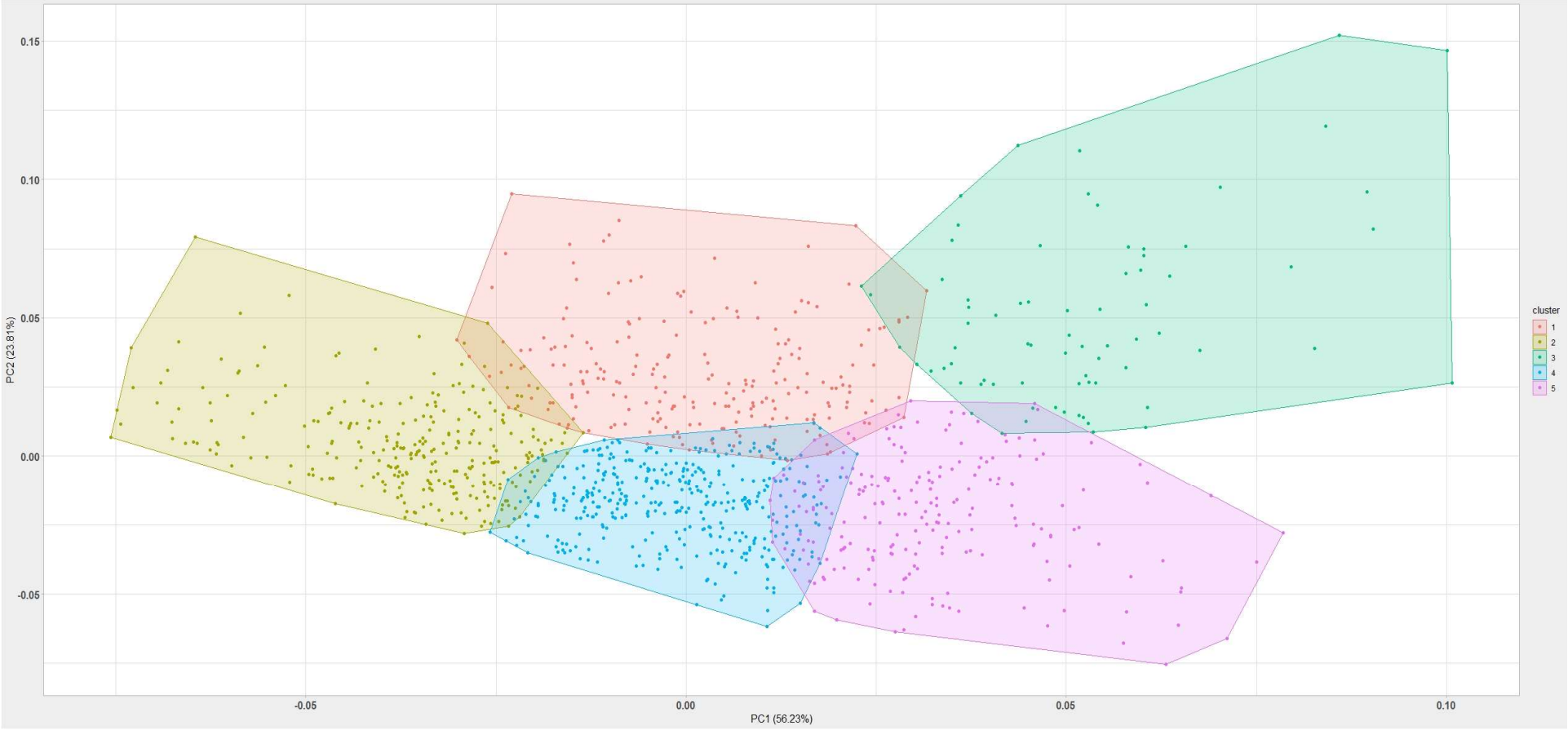
L'individuazione delle serie all'interno del dataset rende possibile calcolarne la durata.

Per calcolare la durata media di uso della vettura, l'algoritmo:

- Individua le serie valide
- Ne calcola la durata
- Somma la durata delle serie valide su base giornaliera
- Calcola la media delle durate giornaliere

Findings

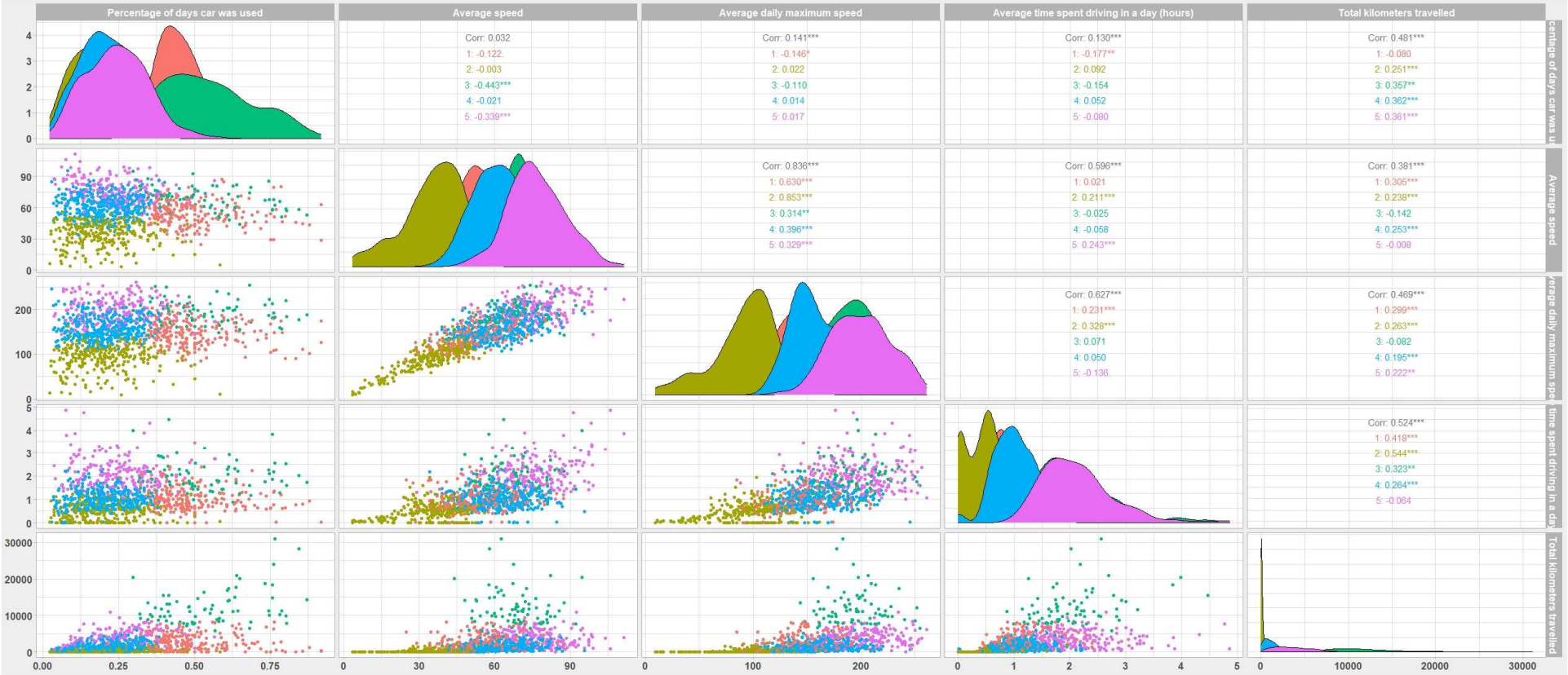
Risultati della clusterizzazione con k-means clustering.



Findings

Descrizione cluster

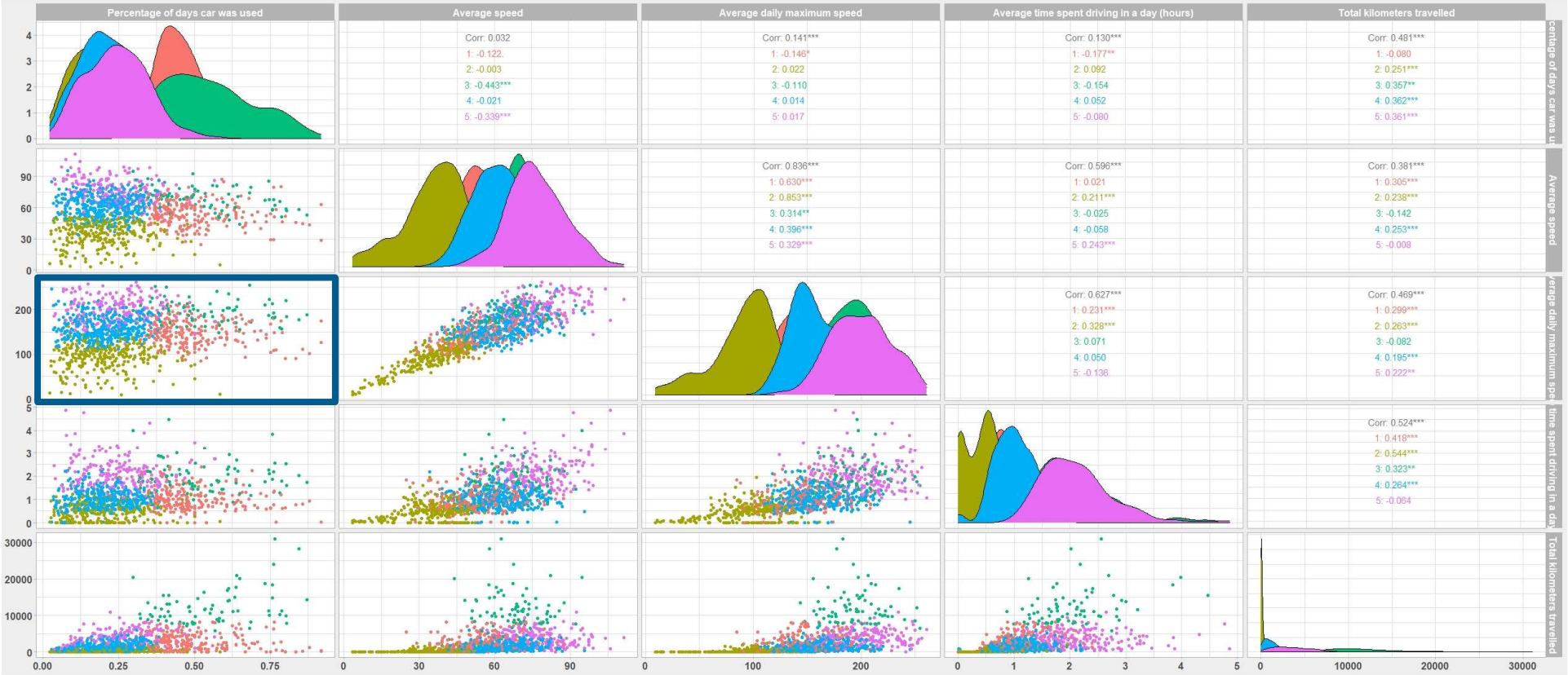
Distribuzioni dei valori delle variabili utilizzate per la clusterizzazione. I colori rappresentano i cluster.



Findings

Descrizione cluster

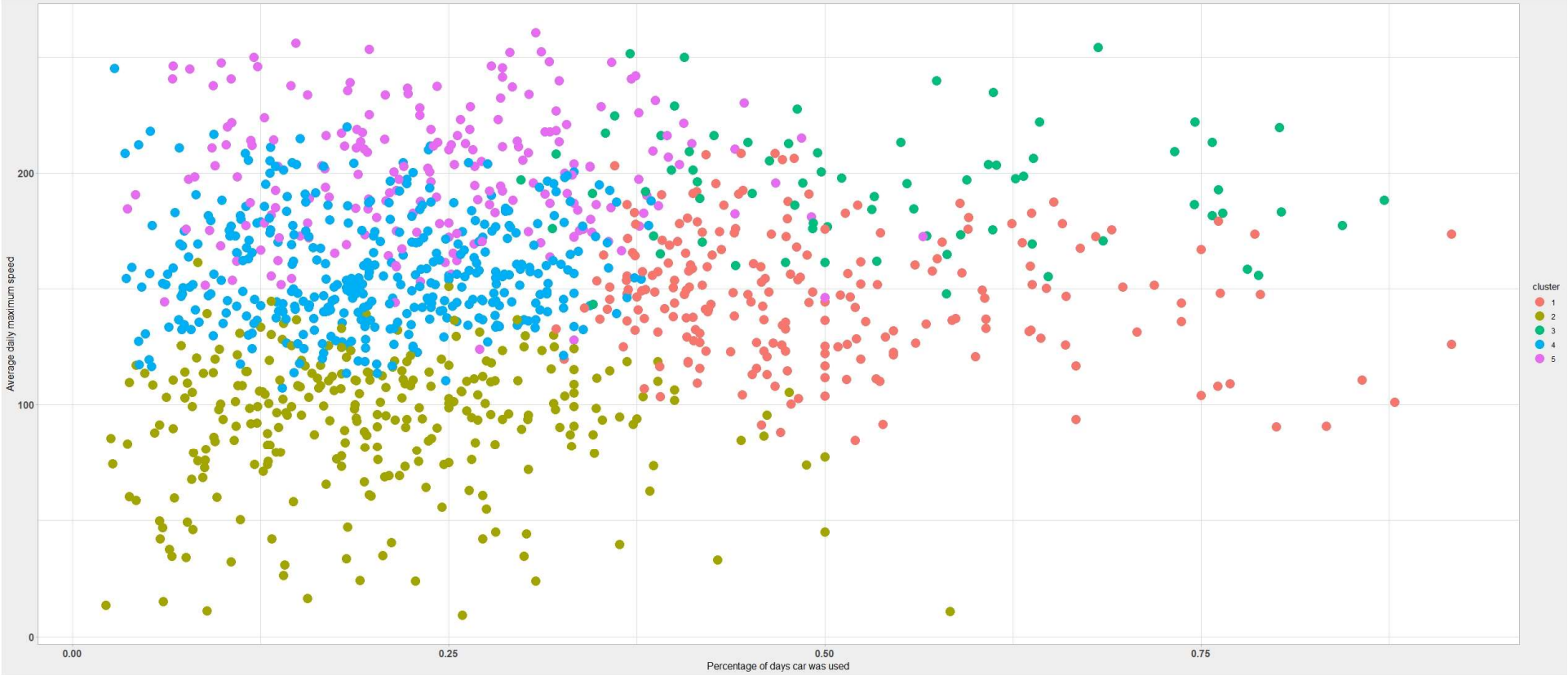
Distribuzioni dei valori delle variabili utilizzate per la clusterizzazione. I colori rappresentano i cluster.



Findings

Descrizione cluster

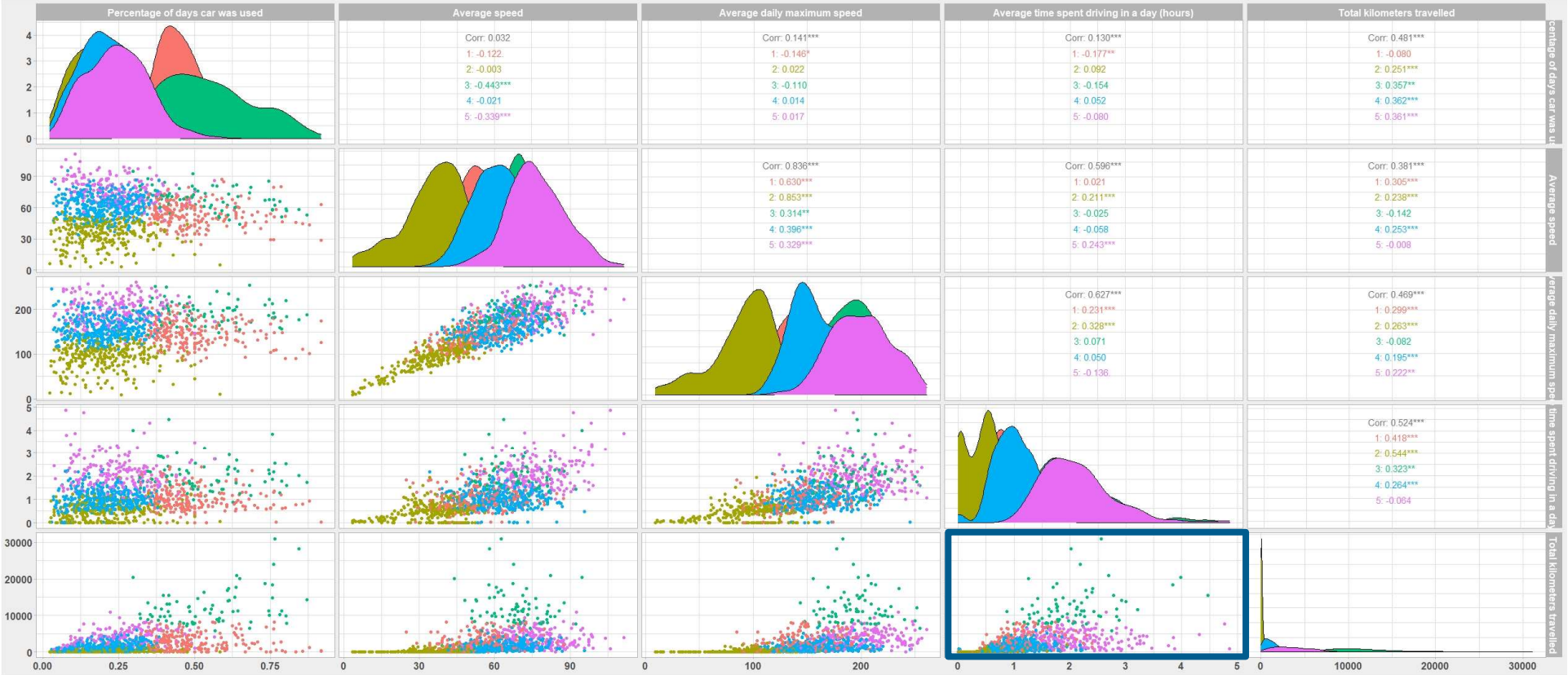
Distribuzione della frequenza di uso dell'auto rispetto alla velocità massima per cluster.



Findings

Descrizione cluster

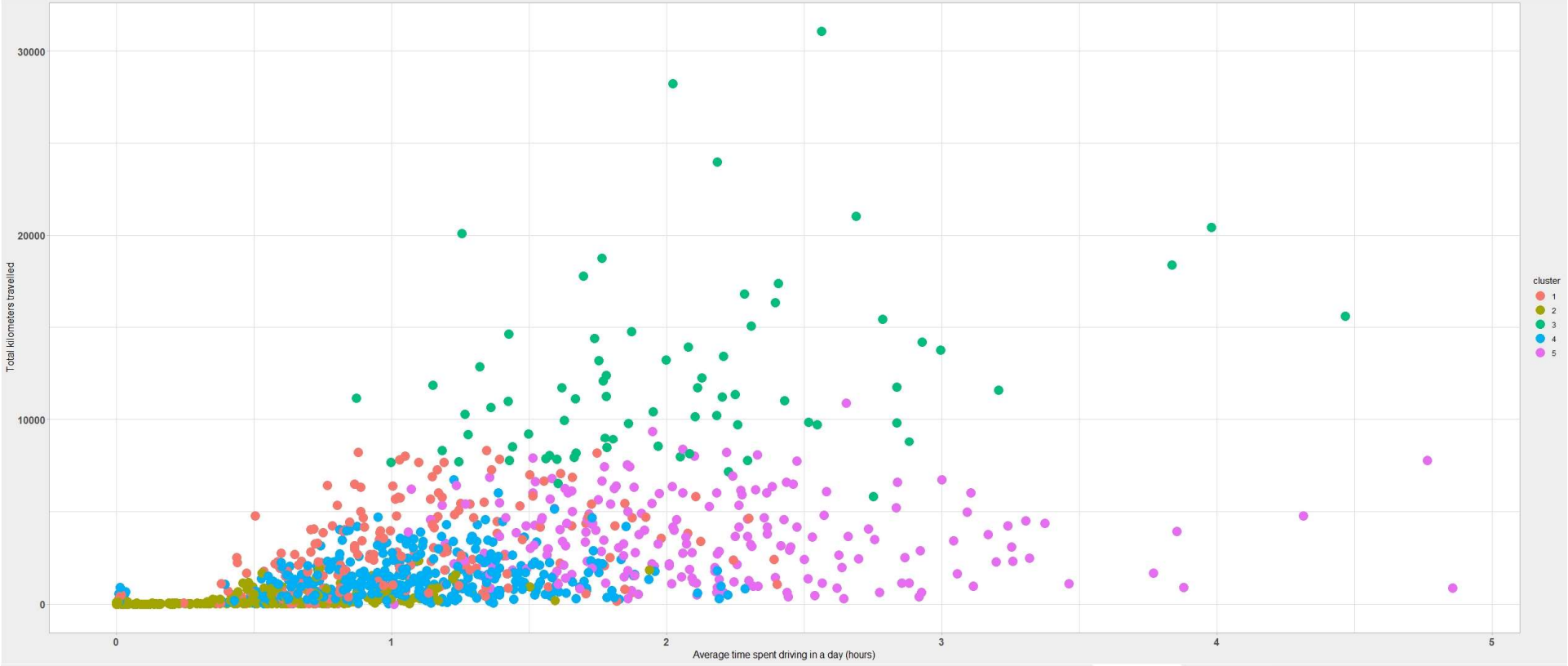
Distribuzioni dei valori delle variabili utilizzate per la clusterizzazione. I colori rappresentano i cluster.



Findings

Descrizione cluster

Distribuzione della durata dell'uso della vettura rispetto ai chilometri percorsi.



Findings

Descrizione cluster

	Velocità massima (km/h)	Velocità media (km/h)	Frequenza utilizzo (%)	Durata uso auto	Distanza totale percorsa (km)	N. vetture
Cluster 1	**147** (85-209)	**54** (25-87)	**50** (32-91)	**1h** (29m-2h24m)	**2611** (27-8272)	221
Cluster 2	**92** (9-161)	**35** (3-56)	**20** (2-58)	**28 min** (0-1h54m)	**249** (1-2356)	285
Cluster 3	**194** (143-254)	**68** (44-95)	**54** (30-87)	**2h** (52m-4h20m)	**12204** (5795-31070)	74
Cluster 4	**158** (107-245)	**61** (34-96)	**20** (2-38)	**1h** (57m-2h17m)	**1351** (36-6683)	391
Cluster 5	**200** (124 – 261)	**77** (47-110)	**24** (3-56)	**2 h** (1h-4h50m)	**3581** (282-10869)	208

In grassetto la media dei valori delle vetture.
Tra parentesi il valore minimo e il valore massimo.

Findings

Descrizione cluster

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa	N. vetture
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve	221
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve	285
Cluster 3	Alta	Alta	Alta	Alta	Lunga	74
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve	391
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve	208

Validazione

- Essendo il clustering un'attività di machine learning non supervisionato, non è possibile verificare matematicamente l'accuratezza dei risultati.

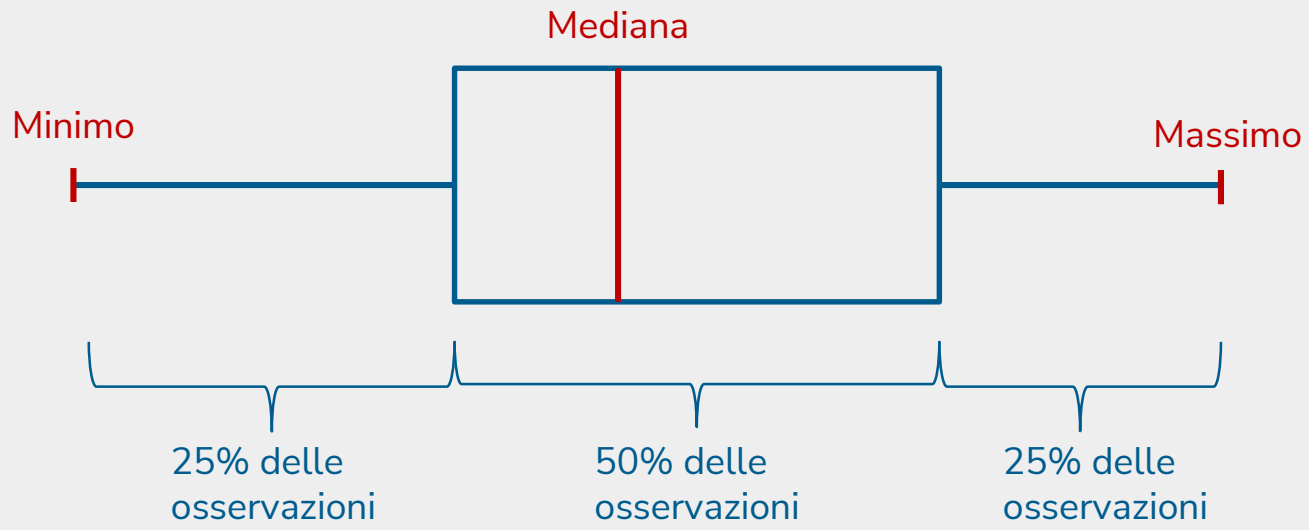
- Occorre quindi adottare una metodologia alternativa per la validazione, basata sul confronto tra i gruppi ottenuti e variabili del dataset che non sono state utilizzate per la clusterizzazione.

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

Validazione

Box plot

- I grafici successivi mostrano delle box plot, tipi di grafici utili a descrivere la distribuzione di dati numerici.

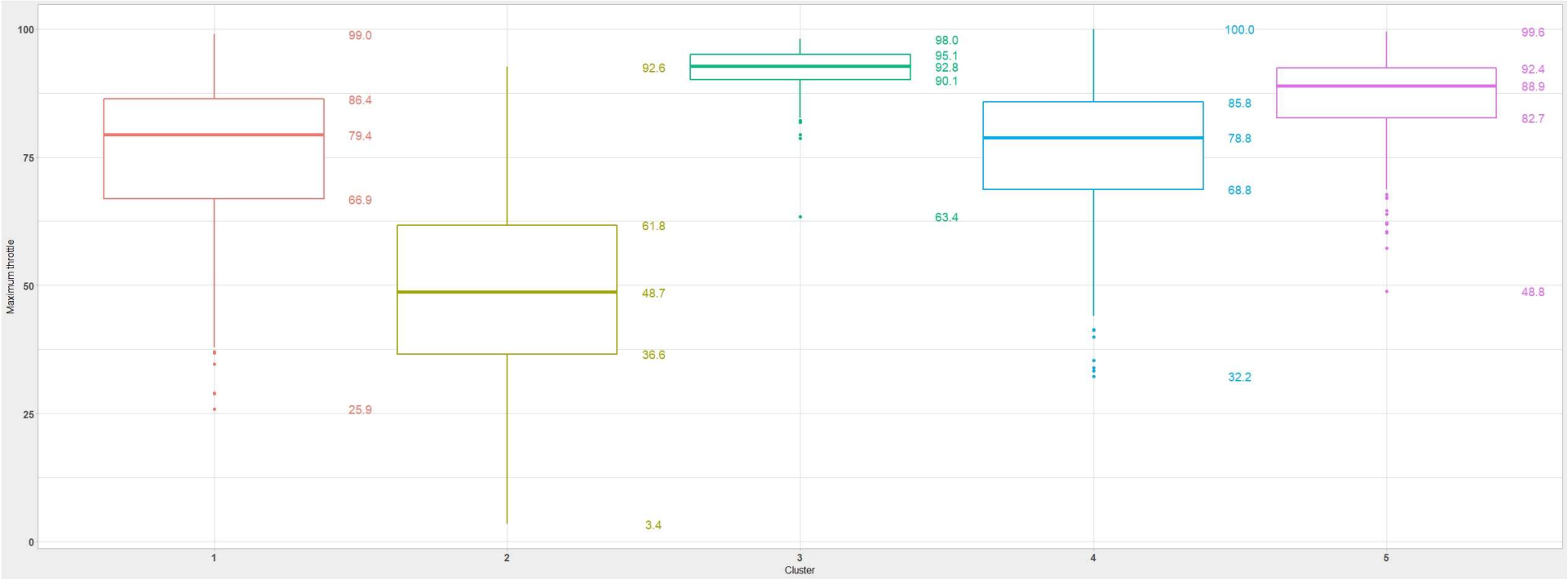


Validazione

Pressione sull’acceleratore

Media dei massimi giornalieri della pressione sull’acceleratore per cluster. La linea orizzontale rappresenta la mediana del cluster. La scatola che la circonda contiene il 50% delle osservazioni. Il restante 50% è distribuito sulle righe verticali.

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

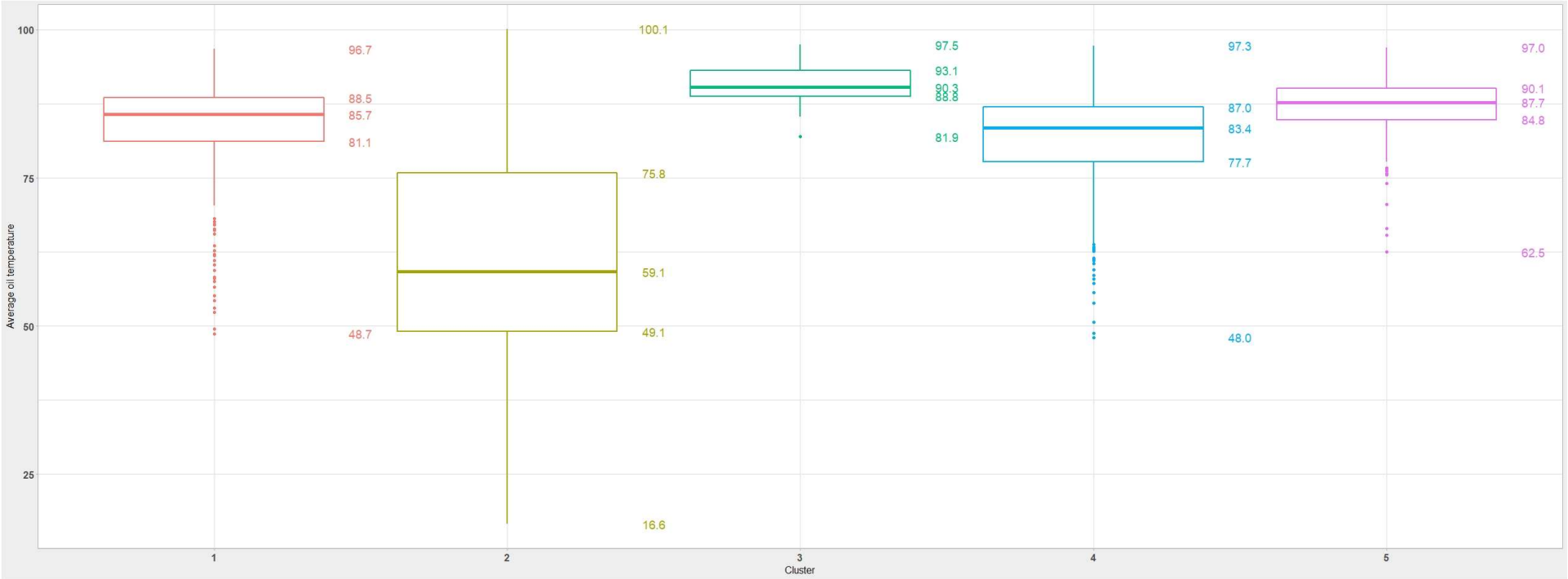


Validazione

Temperatura dell’olio

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Temperatura dell’olio media per cluster.

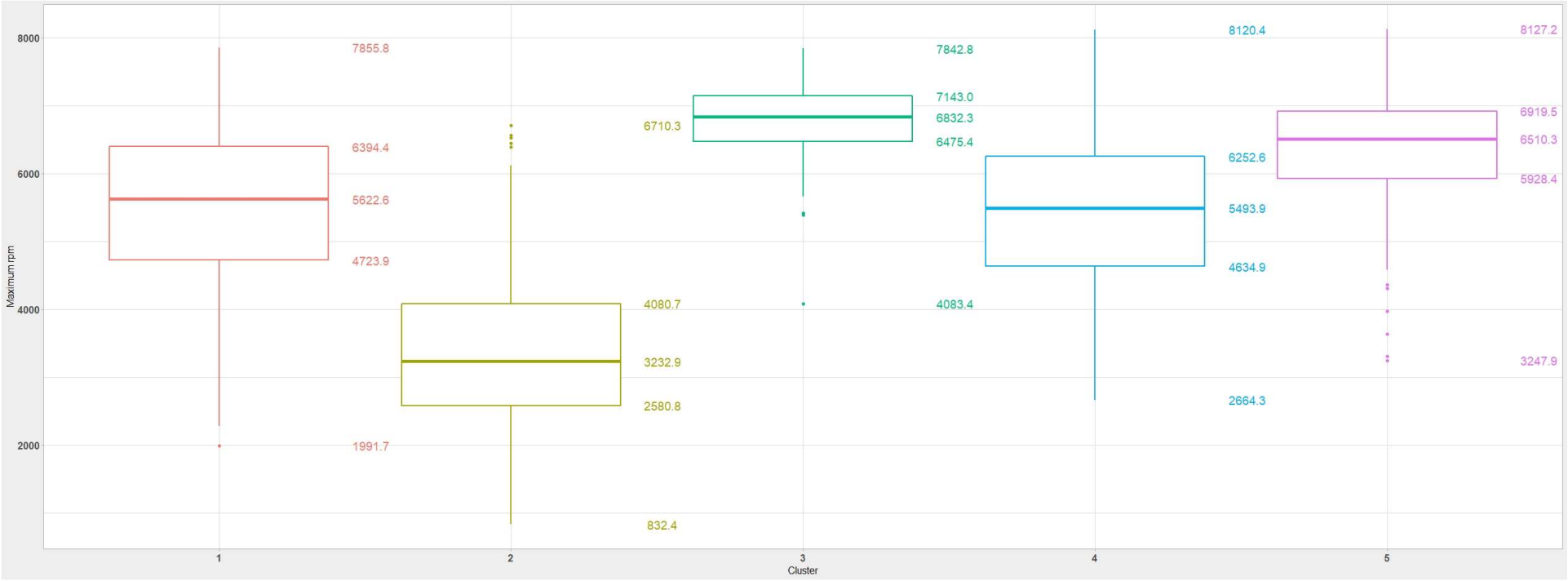


Validazione

Rotazioni per minuto

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Media dei massimi giornalieri delle rotazioni per minuto per cluster.

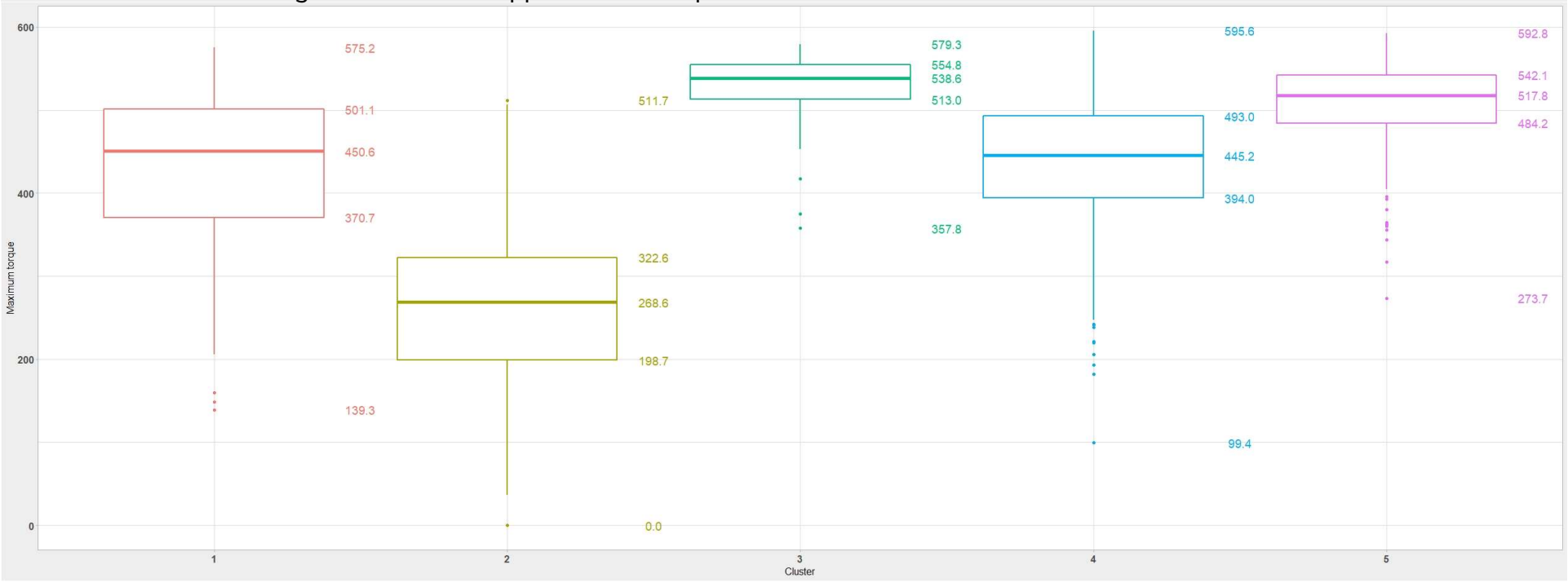


Validazione

Coppia istantanea

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Media dei massimi giornalieri della coppia istantanea per cluster.

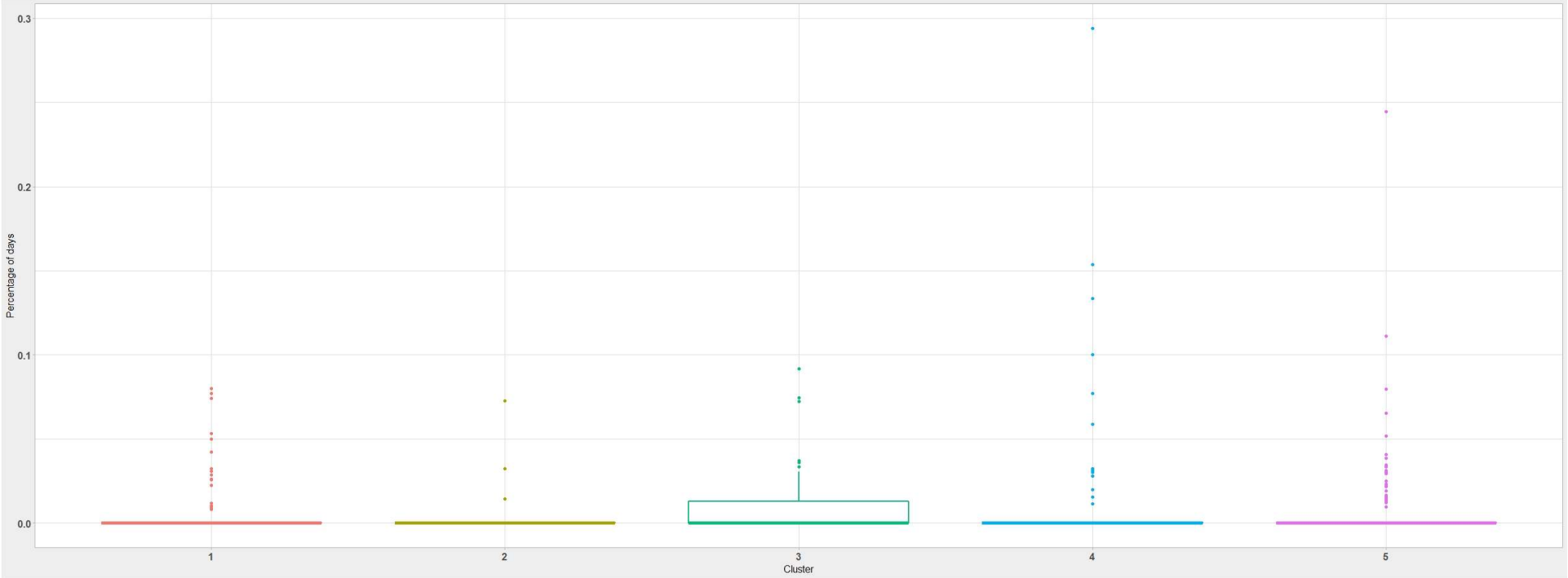


Validazione

Active launch control

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

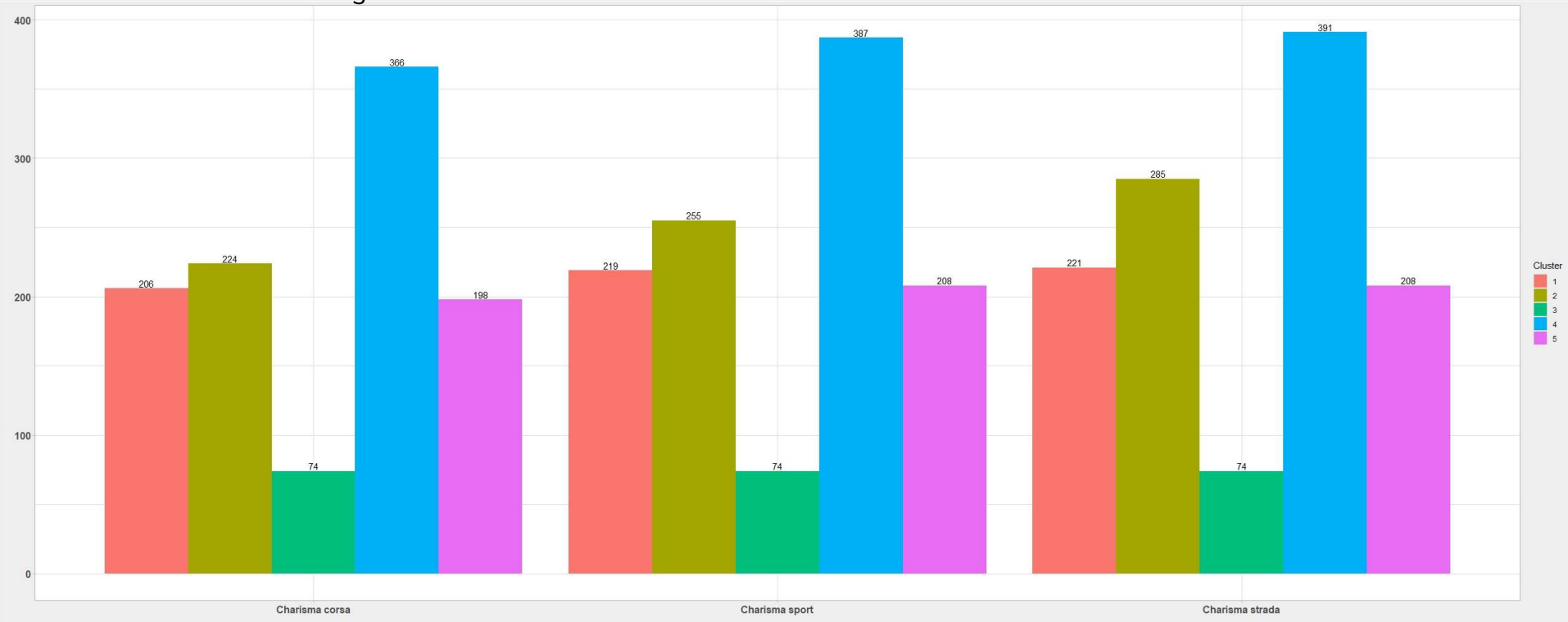
Percentuale di giorni sul totale in cui l'active launch control è stato attivato almeno una volta.



Validazione
Driving mode

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa	N. vetture
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve	**221**
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve	**285**
Cluster 3	Alta	Alta	Alta	Alta	Lunga	**74**
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve	**391**
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve	**208**

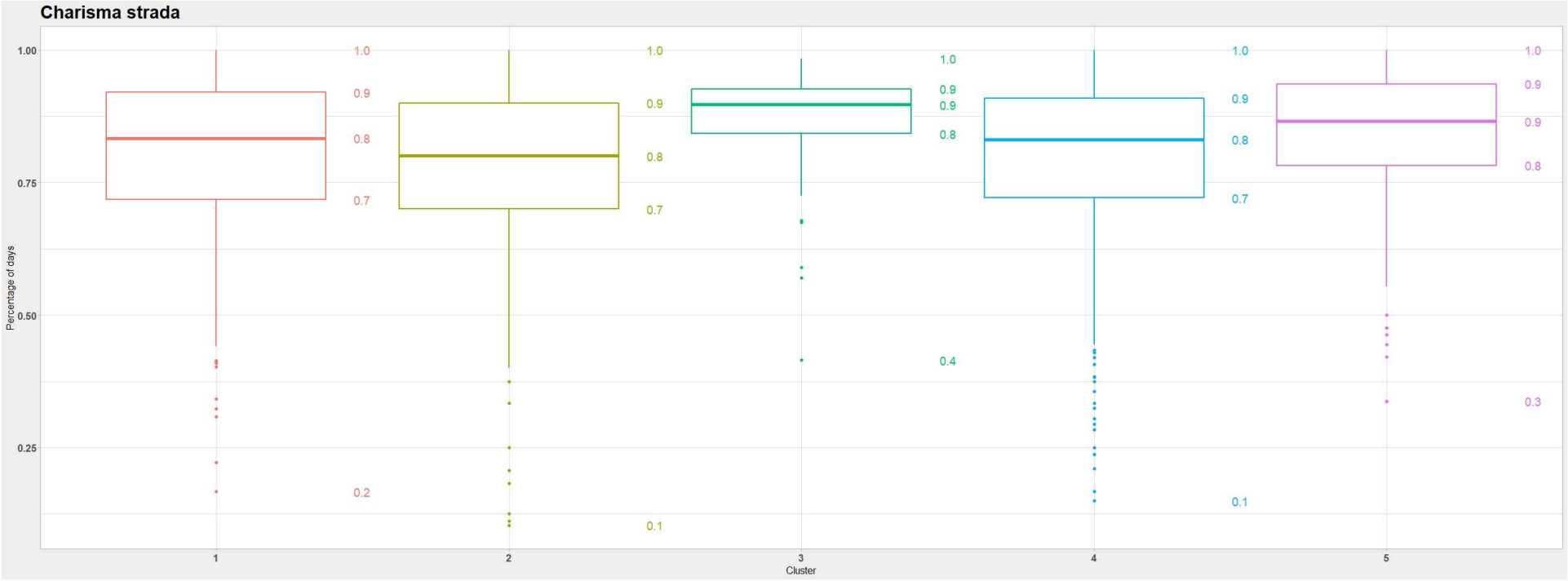
Numero di auto che ha utilizzato almeno una volta la modalità di guida.



Validazione
Driving mode

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

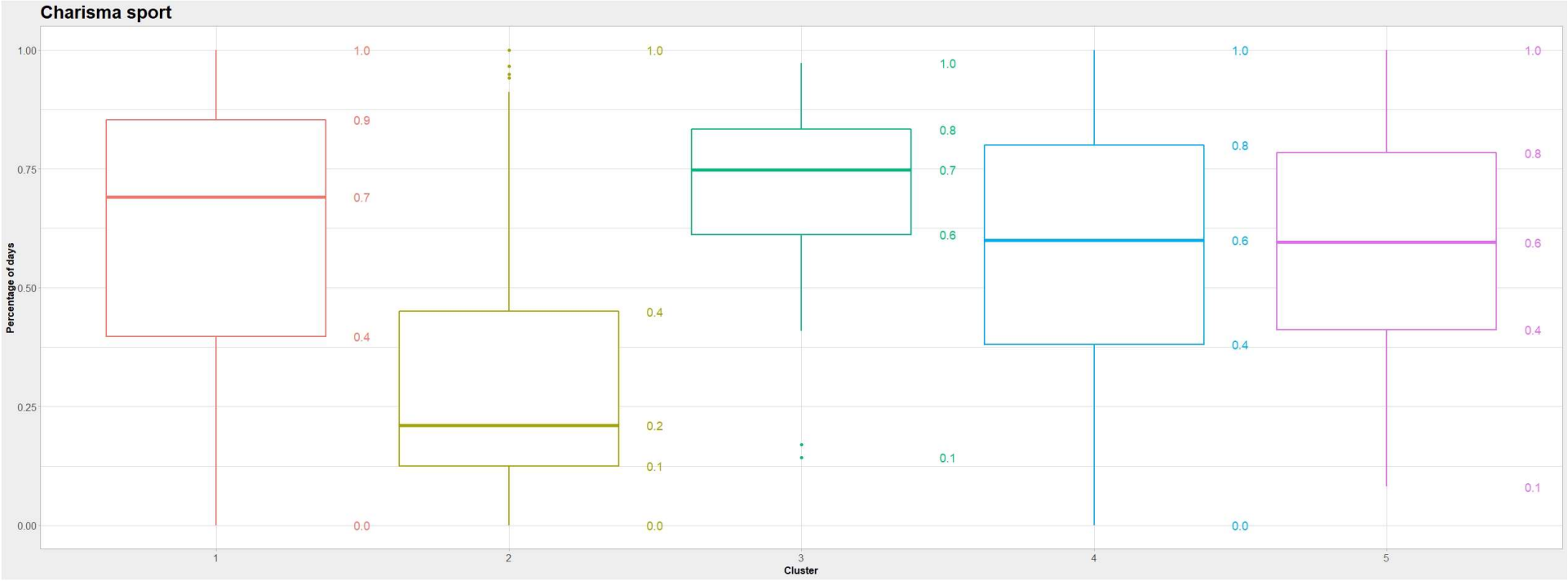
Percentuale di giorni sul totale in cui è stata utilizzata almeno una volta la modalità di guida 1 – **Charisma strada**.



Validazione
Driving mode

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

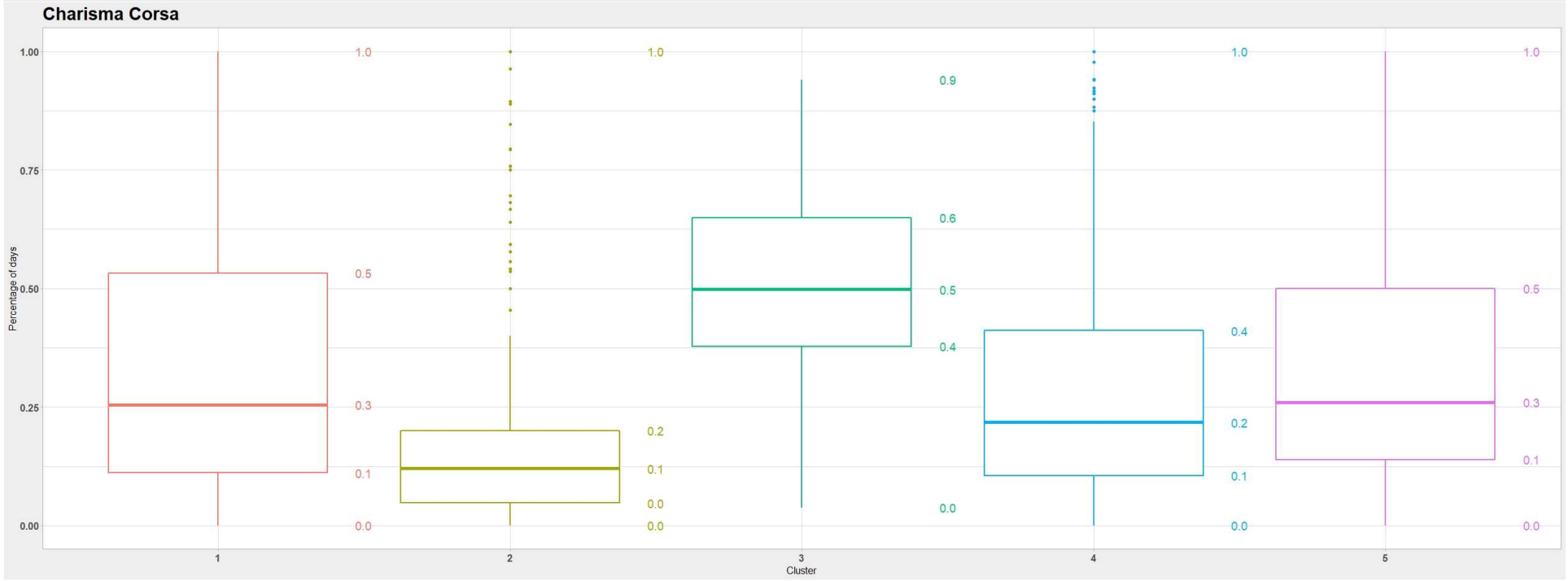
Percentuale di giorni sul totale in cui è stata utilizzata almeno una volta la modalità di guida 2 – **Charisma sport**.



Validazione
Driving mode

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Percentuale di giorni sul totale in cui è stata utilizzata almeno una volta la modalità di guida 3 – **Charisma corsa**.



applied

innovation makers

Possibili applicazioni

La metodologia e gli strumenti sviluppati nel corso dell'analisi hanno numerose potenziali applicazioni in ulteriori aree:

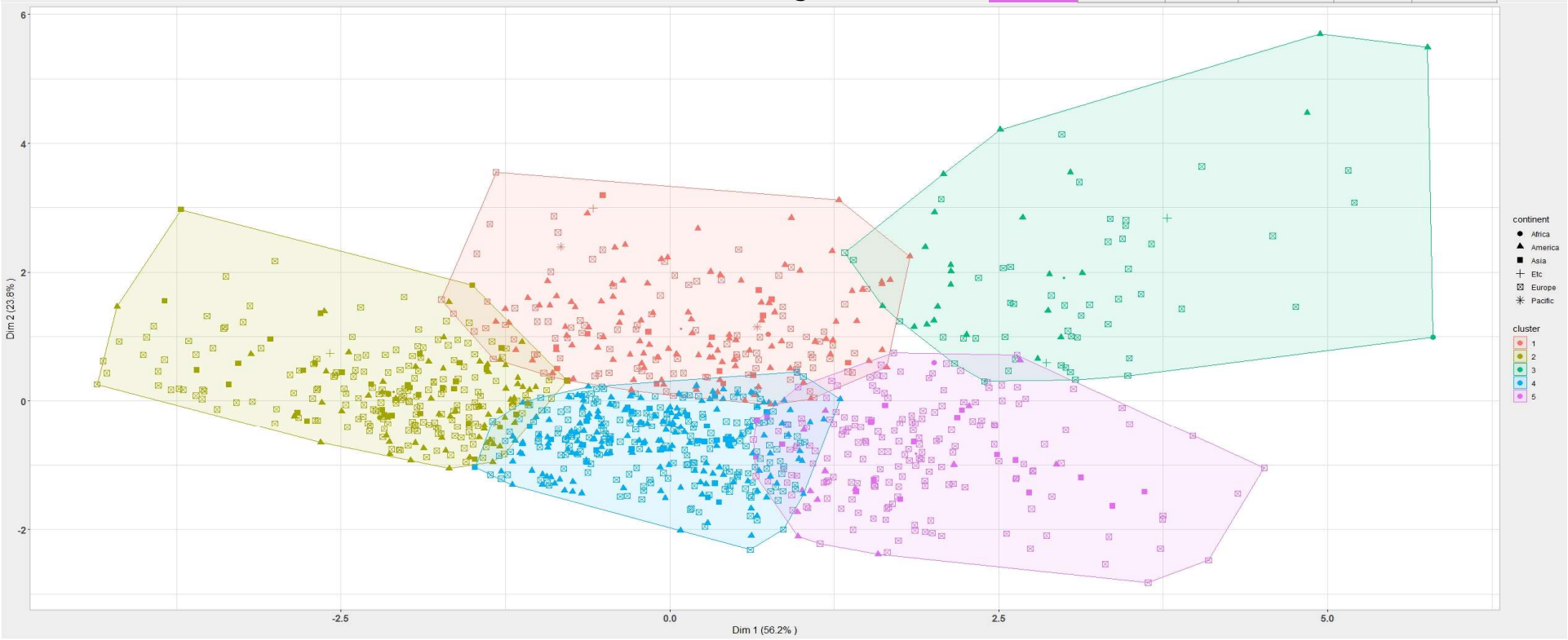
- Utilizzo del dato per determinare le condizioni della vettura al fine di prevedere la necessità di manutenzione (predictive maintenance)
- Assistenza nella gestione del dato nei sistemi Lamborghini in modo da renderlo più agevolmente utilizzabile per future analisi e incorporamento degli algoritmi sviluppati nei sistemi Lamborghini
- **Sviluppo ulteriore della categorizzazione su base geografica**
- **Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore**
- **Individuazione di periodi trascorsi in pista tramite pattern recognition**

Possibili applicazioni

Sviluppo della categorizzazione su base geografica

Risultati della clusterizzazione con k-means clustering.

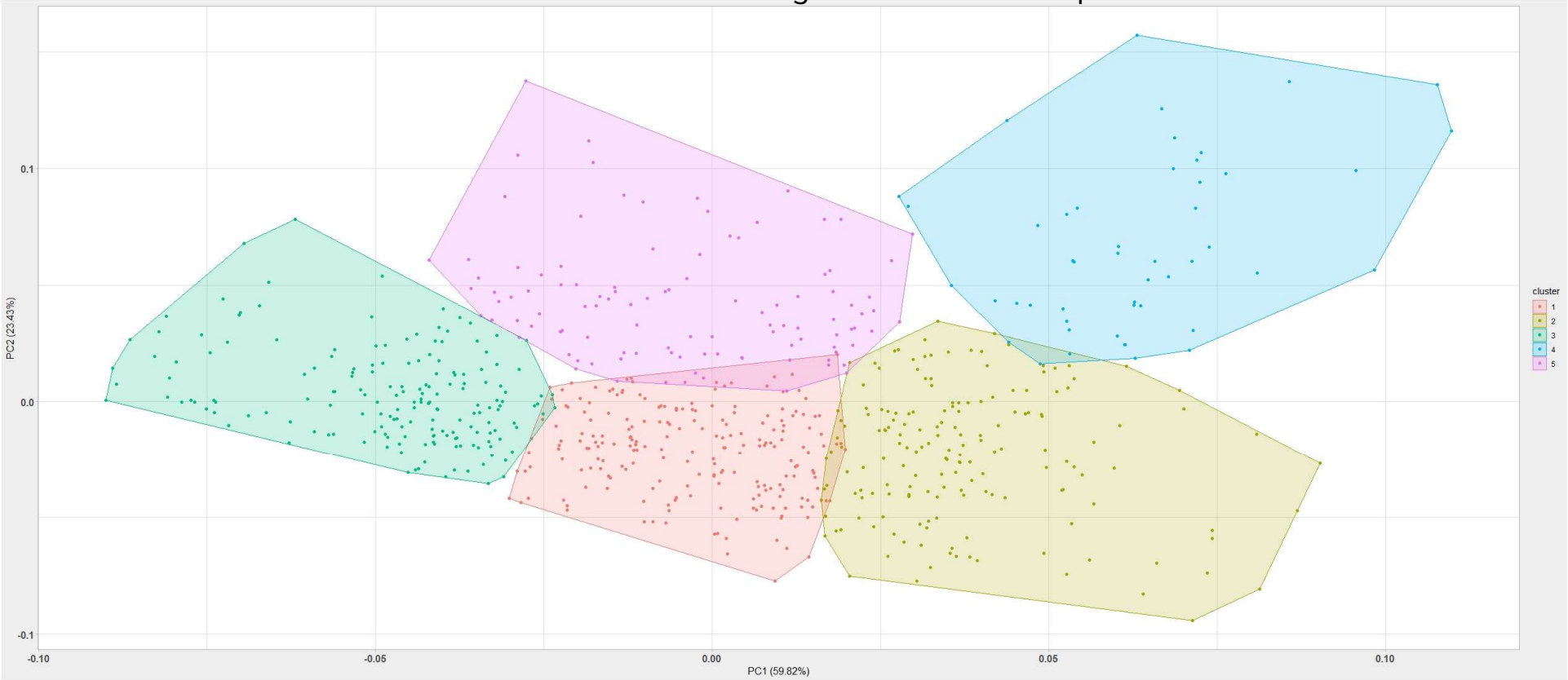
	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve



Possibili applicazioni

Sviluppo della categorizzazione su base geografica

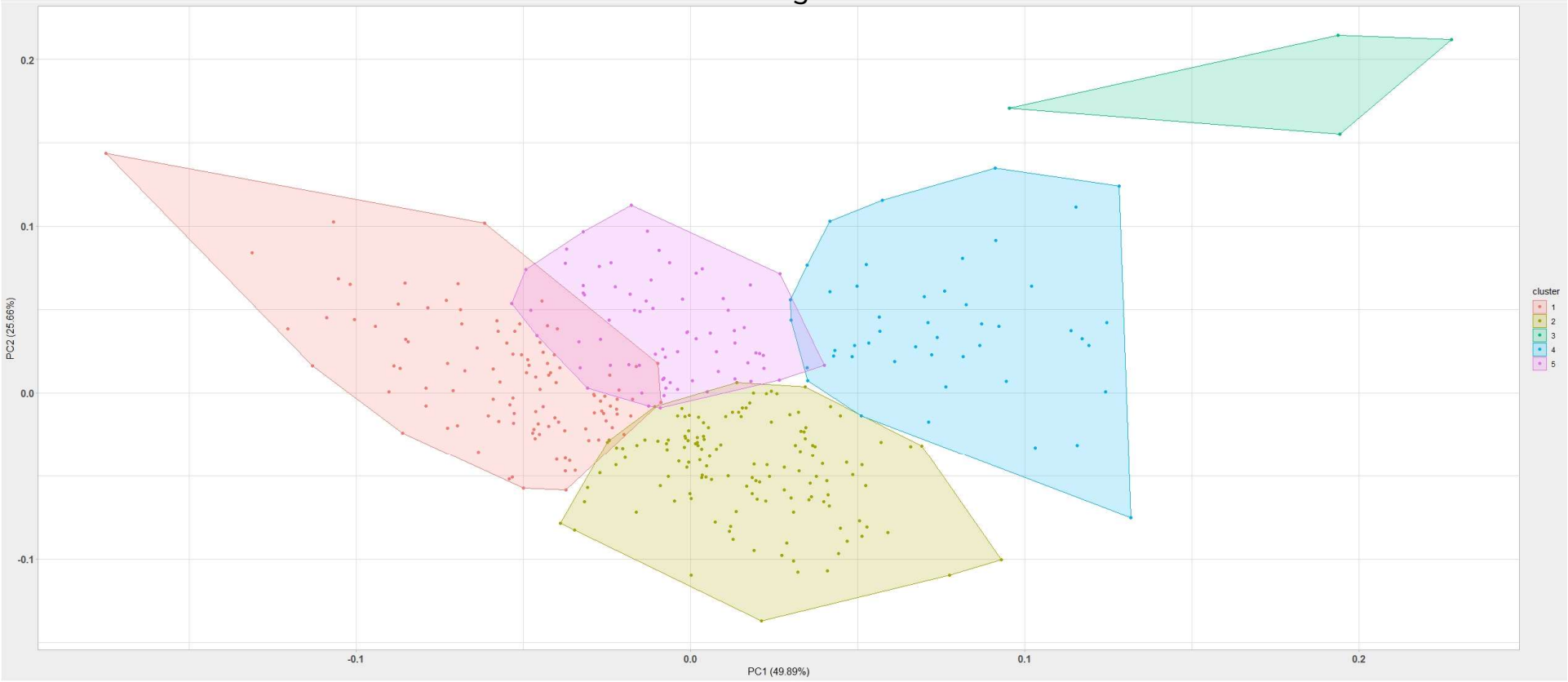
Risultati della clusterizzazione con k-means clustering di **709** auto in Europa.



Possibili applicazioni

Sviluppo della categorizzazione su base geografica

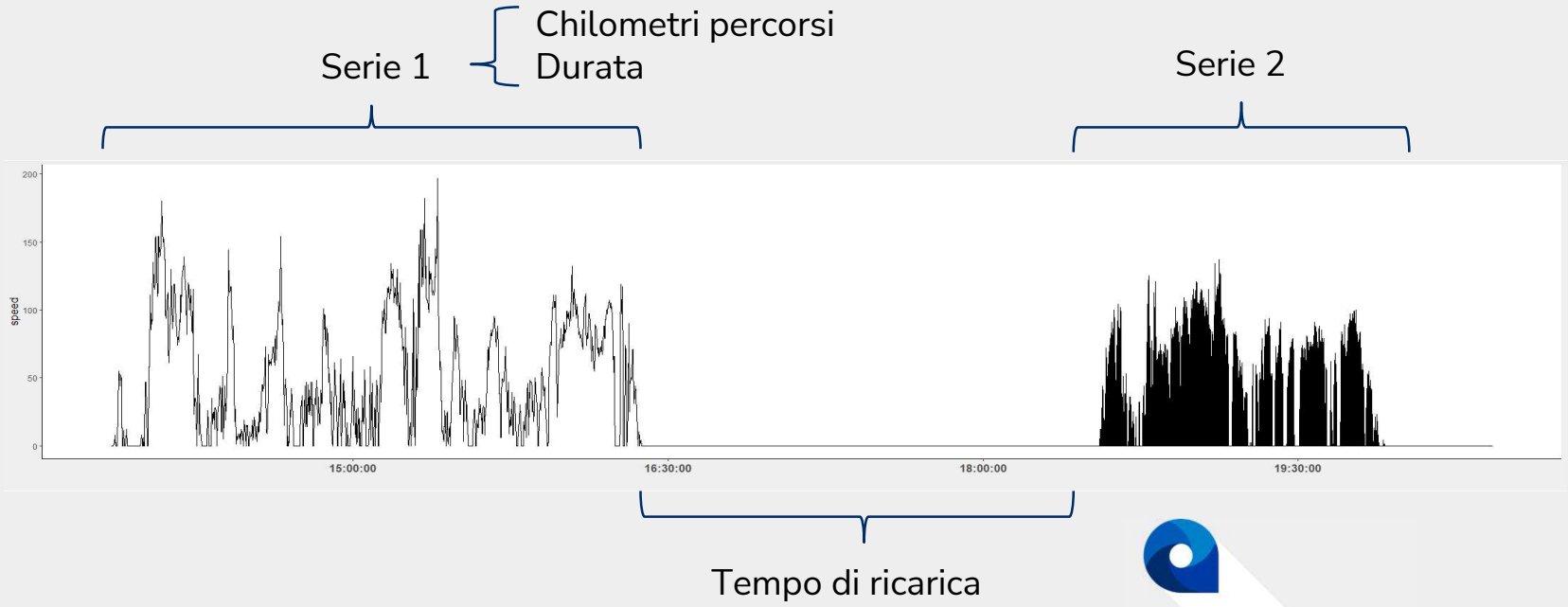
Risultati della clusterizzazione con k-means clustering di **368** auto in America.



Possibili applicazioni

Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore

- I parametri dell'algoritmo della valorizzazione delle serie possono essere impostati in modo da considerare come serie distinte solo quelle tra cui intercorre un sufficiente tempo di ricarica

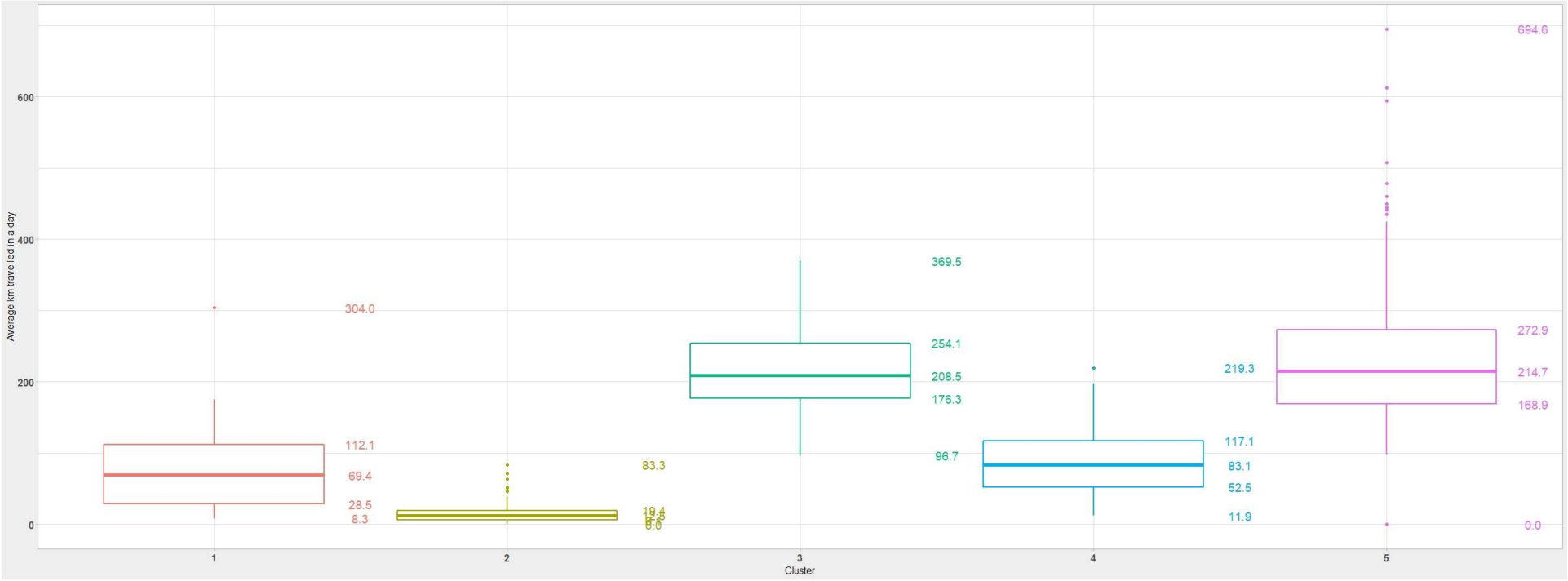


Possibili applicazioni

Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Km percorsi in media al giorno per cluster.

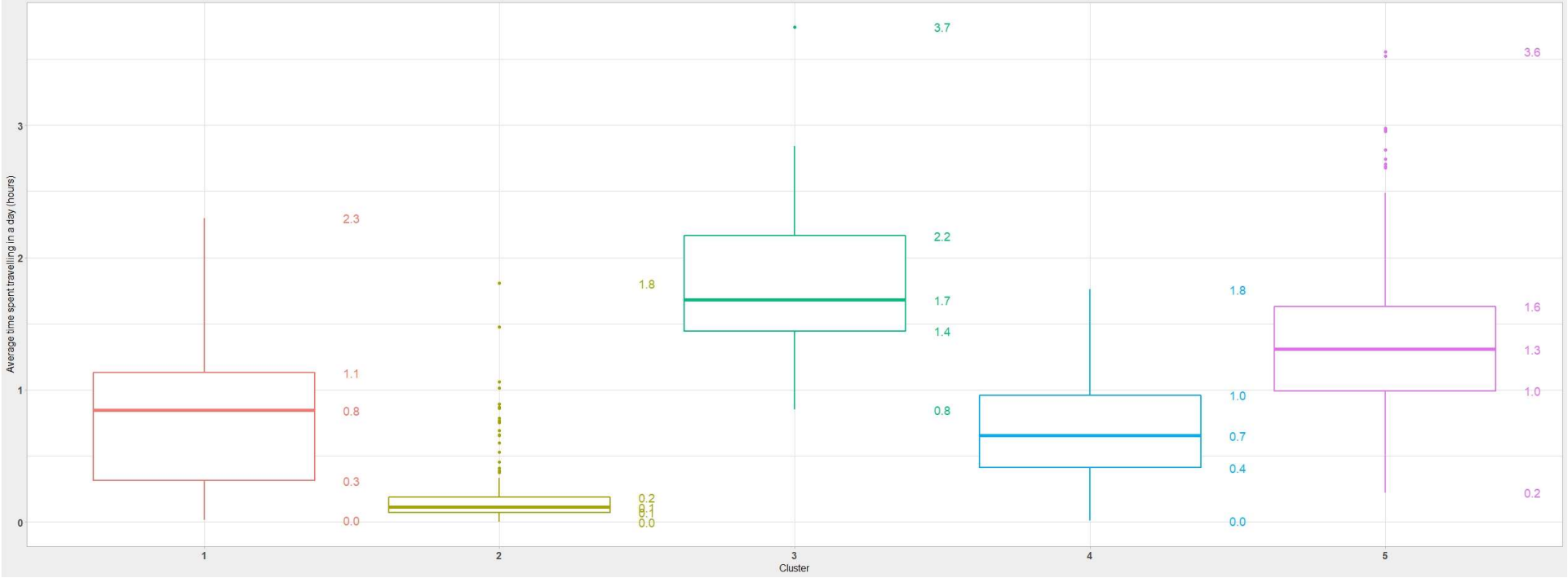


Possibili applicazioni

Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Durata di guida giornaliera media per cluster.

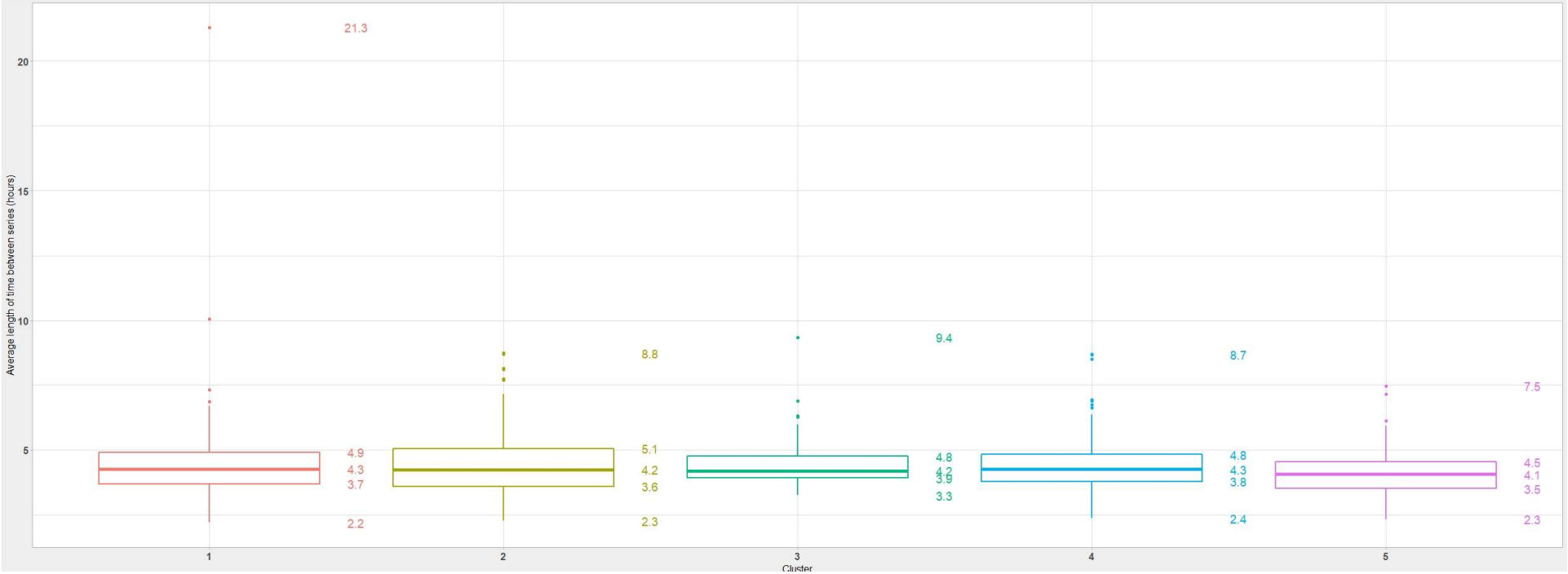


Possibili applicazioni

Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Tempo di ricarica: distanza media giornaliera tra serie (ore).

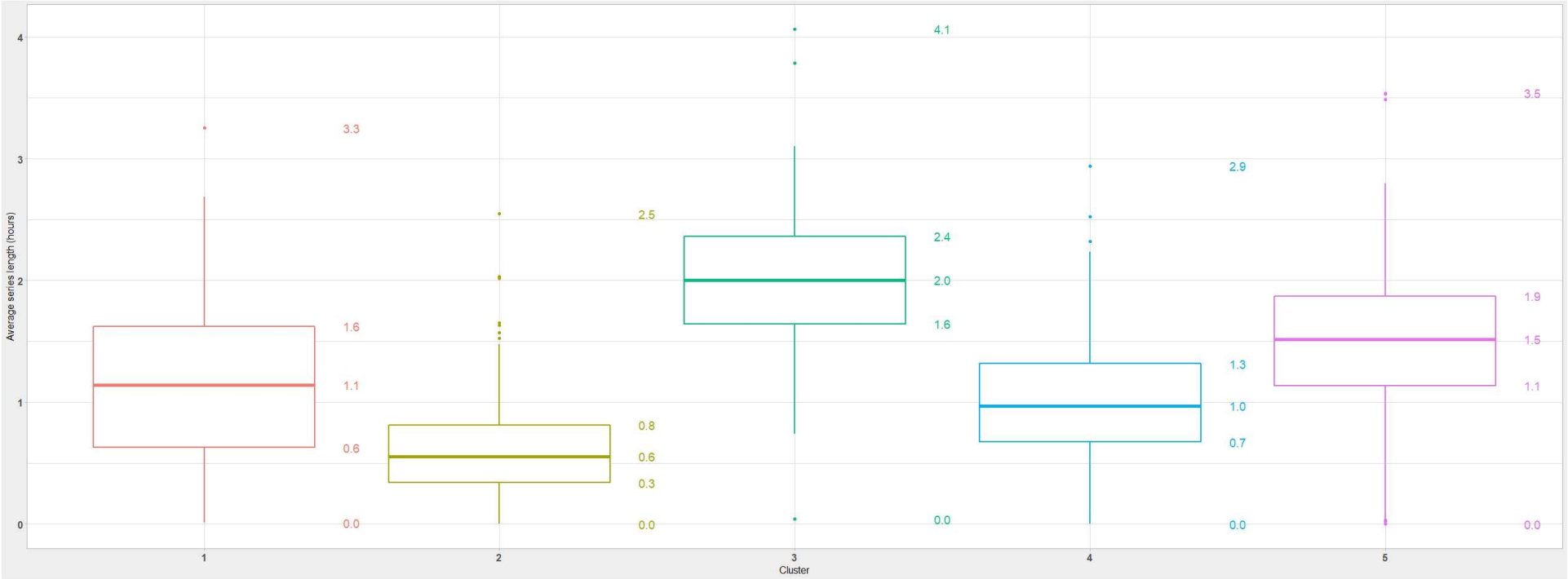


Possibili applicazioni

Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Durata media delle serie per cluster (ore).

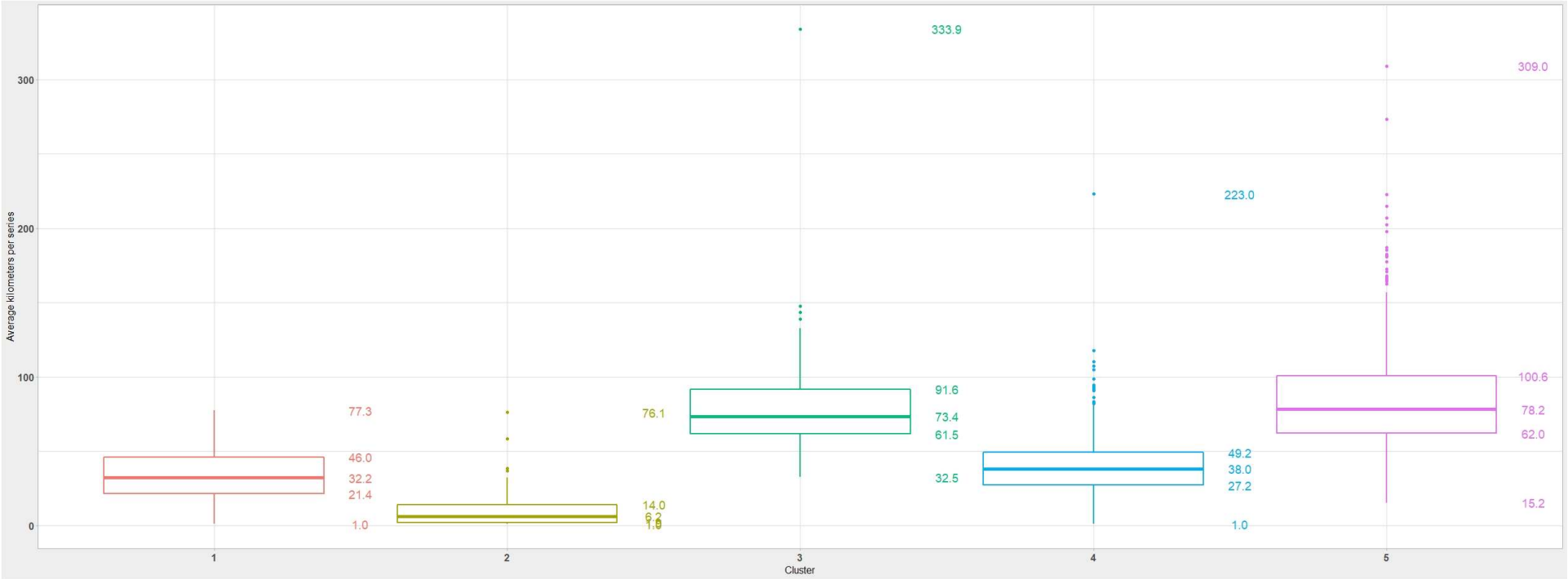


Possibili applicazioni

Utilizzo del dato per indirizzare futuri progetti di elettrificazione del motore

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Km medi percorsi in una serie per cluster.

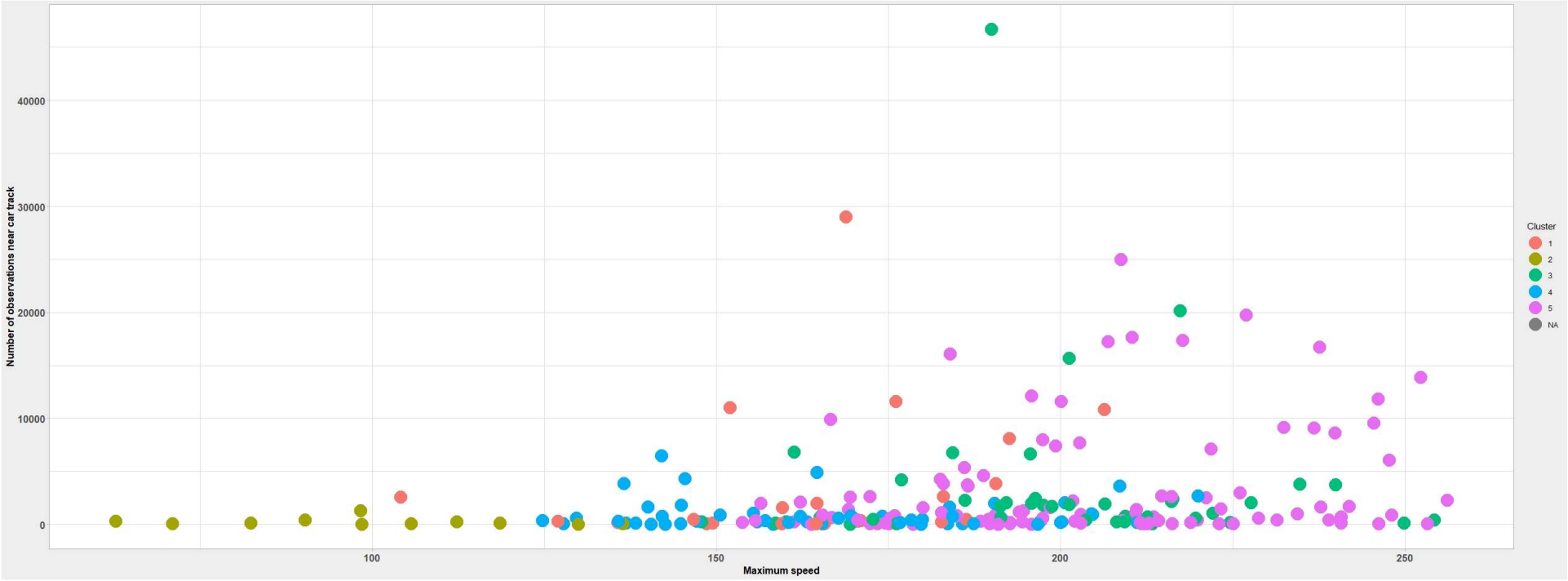


Possibili applicazioni

Individuazione periodi di guida in pista

	Velocità massima	Velocità media	Frequenza utilizzo	Durata uso auto	Distanza totale percorsa
Cluster 1	Media	Media	Alta	Medio-bassa	Medio-breve
Cluster 2	Bassa	Bassa	Medio-bassa	Bassa	Breve
Cluster 3	Alta	Alta	Alta	Alta	Lunga
Cluster 4	Media	Media	Medio-bassa	Medio-bassa	Breve
Cluster 5	Alta	Alta	Medio-bassa	Alta	Medio-breve

Numero di osservazioni per vettura rilevate nei pressi di 60 dei maggiori autodromi europei, confrontate con le velocità massime dei veicoli.



applied

innovation makers

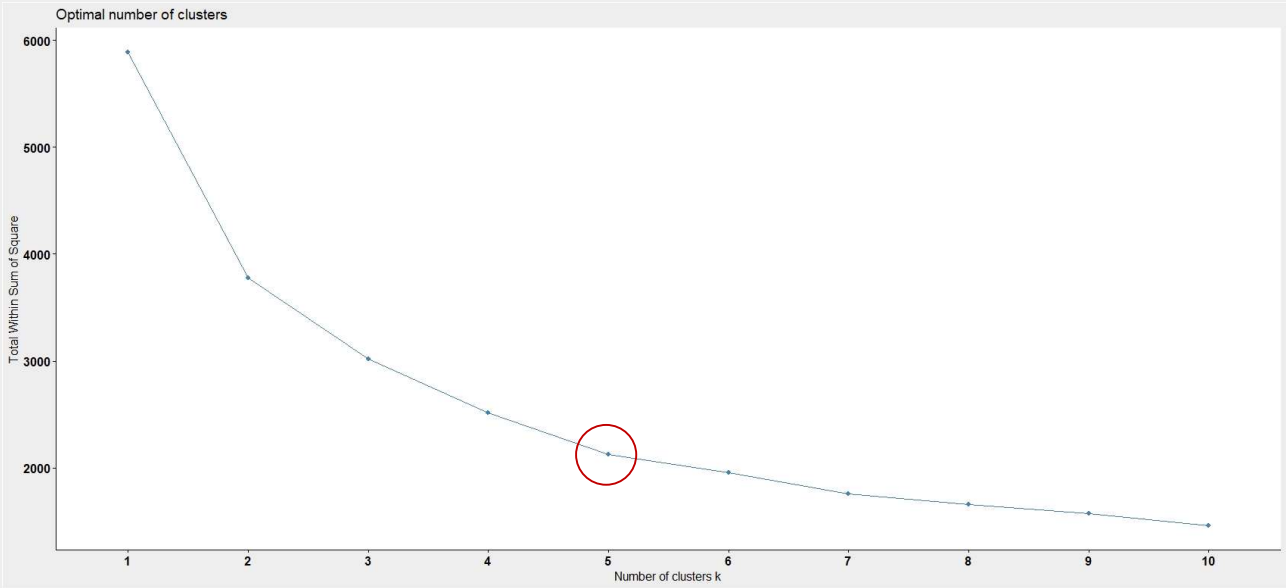
Annex

Numero di cluster

Esistono una varietà di tecniche empiriche per la selezione di k:

1. **Metodo del gomito:**

   Si itera l'algoritmo per diversi valori di k, per ognuno si calcola la somma delle distanza al quadrato tra ogni centroide ed i punti del relativo cluster.



Annex

Numero di cluster

Esistono una varietà di tecniche empiriche per la selezione di k:

2. **Gap statistics:**

   Si compara la variazione totale interna al cluster per vari possibili valori di k 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 è maggiore.