UNIVERSITÀ
POLITECNICA
DELLE MARCHE

FACOLTA' DI INGEGNERIA

CORSO DI LAUREA IN INGEGNERIA GESTIONALE

Analisi delle traiettorie dei consumatori in ambito retail

Analysis of customers' trajectories in retail environment

Candidato:

Federica IEZZI

Relatore:

Prof. Primo ZINGARETTI

Correlatore:

Dott.ssa Marina PAOLANTI

Anno accademico 2020/2021

Alla mia famiglia, che mi ha supportato in questo percorso

SOMMARIO

In questo elaborato si propone un'analisi, mediante un algoritmo di clustering gerarchico, di un dataset relativo a traiettorie di consumatori in uno spazio chiuso. Alla quale segue una fase di valutazione della qualità dei risultati ottenuti dall'algoritmo tramite una verifica di dissimilarità, ed infine si procede con un'analisi del comportamento dei consumatori con la quale si vuole comprendere meglio la condotta di quest'ultimi in ambiente retail.

Indice

1. INTRODUZIONE

Negli ultimi anni, l'attenzione degli studiosi si è spesso focalizzata sulle abitudini di acquisto dei consumatori, poiché in un ambiente commerciale sempre più competitivo riuscire a comprendere i bisogni di quest'ultimi determina un grande vantaggio per le aziende; in quanto una conoscenza dettagliata del comportamento dei consumatori permette di adottare soluzioni efficienti nel soddisfacimento delle necessità dei propri clienti, portando l'azienda ad avere dei buoni risultati nelle vendite e conseguentemente ad ottenere un'influenza positiva sulla profittabilità.

Questo è il motivo per cui molti imprenditori, oggigiorno, sono interessati a adottare un approccio più centrato sul consumatore.
Quindi, si cercano metodologie sempre più innovative ed efficienti per collezionare dati e successivamente analizzarli, in modo da estrapolarne informazioni che le aziende possano utilizzare per implementare un sistema di supporto alle decisioni, fine alla massimizzazione dell'impatto che i propri prodotti o servizi hanno sulle persone.
Un altro fine commerciale che è possibile ricavare dall'analisi dei dati è lo studio e l'individuazione della presenza di schemi nel comportamento dei consumatori, i quali possono essere utilizzati per individuare le caratteristiche di ciascuna categoria di utenti, in modo da progettare e offrire servizi specifici e proporre quindi esperienze d'acquisto ad hoc per ogni cluster di clienti che è stato individuato successivamente all'analisi del dataset utilizzato. Per questa ragione, le aziende si trovano a dover fronteggiare problemi pragmatici nel realizzare quanto necessario per la diversificazione della gamma di beni o servizi, trovandosi costretti ad ottimizzare la loro pianificazione, le proprie attività promozionali e le attività inerenti alla produzione e alla distribuzione.

In questo elaborato si tratta di ambiente retail, quindi si presuppone la presenza fisica delle persone all'interno del negozio, necessaria per la raccolta dei dati, come ad esempio gli spostamenti all'interno del locale oppure i prodotti davanti ai quali più persone sostano, però nell'ultimo decennio, con una diffusione sempre più capillare e una alfabetizzazione digitale spinta soprattutto dalle condizioni storiche attuali, abbiamo visto una crescita esponenziale di attività commerciali online. Questo argomento esula dalla trattazione di questa tesi ma è degno di nota nell'ambito della raccolta dati e dello studio del comportamento d'acquisto, in quanto mediante l'uso di concetti come quello dei cookie e specialmente se si acquista metodicamente dagli stessi siti è molto più semplice raccogliere dati sulle abitudini d'acquisto e sulle preferenze personali dei consumatori e conseguentemente essere sottoposti a pubblicità mirate agli interessi della persona.

La raccolta dei dati in ambiente retail, invece, deve essere effettuata in modo implicito, ovvero la tecnologia presente in loco per la rilevazione di attività non deve essere notata da colui che si trova a scegliere un prodotto o una direzione perché potrebbe determinare un'alterazione delle intenzioni iniziali della persona che sarebbe così influenzata dalla presenza delle macchine e si andrebbe ad alterare il normale comportamento d'acquisto, rendendo inutilizzabili i dati per uno studio che ha come scopo finale quello di analizzare la normale condotta dei clienti all'interno di un negozio. Per cui, onde evitare situazioni simili, si è adottata una tecnologia che rende l'esperienza d'acquisto accogliente per la presenza umana [1,2], ovvero un ambiente in cui gli utenti sono circondati da interfacce intelligenti intuitive inserite in qualsiasi tipo di oggetto, in grado di riconoscere e rispondere alla presenza di diversi individui. Infatti, negli ambienti retail intelligenti vengono efficientemente utilizzati sistemi integrati per monitorare i clienti e studiarne il comportamento e l'interattività con l'ambiente che li circonda e con le altre persone presenti nel negozio, mettendo in gioco tecnologie economiche per questo tipo di applicazione [3,4]. I dati vengono raccolti utilizzando sensori installati nei vari locali con lo scopo di monitorare:

- il livello di attrazione che le persone hanno nei confronti del negozio, ad esempio quante persone si avvicinano e decidono di entrare;
- l'attenzione che un determinato marchio è in grado di attirare, tenendo conto del tempo che un individuo sosta di fronte un determinato scaffale;
- le azioni che i clienti compiono, notando il modo in cui questi interagiscono con i prodotti presenti all'interno del negozio;

Queste variabili sono importanti, poiché ci danno delle informazioni sull'impatto che determinati prodotti hanno sulle persone, se la loro esposizione è sufficiente, oppure se la pubblicità che è stata programmata è stata efficace per promuovere quel determinato bene, in aggiunta a ciò, ci indicano anche se un prodotto si trova in una posizione che permette esso di catturare e mantenere l'attenzione degli individui passanti.

Lo scopo di questo elaborato è quello di trarre informazioni fini al miglioramento dell'esperienza d'acquisto dei consumatori effettuando un clustering sulle traiettorie degli equipaggiamenti d'acquisto (carrelli/cestini) in un ambiente di vendita al dettaglio durante le ore di apertura dell'attività.

Il clustering (dal termine inglese cluster analysis introdotto da Robert Tryon nel 1939) è un insieme di tecniche di analisi multivariata dei dati volte alla selezione e raggruppamento di elementi omogenei in un insieme di dati. Le tecniche di clustering si basano su misure relative alla somiglianza tra gli elementi. In molti approcci questa similarità, o meglio, dissimilarità, è concepita in termini di distanza in uno spazio multidimensionale. La qualità delle analisi ottenute dagli algoritmi di clustering dipende molto dalla scelta della metrica, e quindi da come è calcolata la distanza. Gli algoritmi di clustering raggruppano gli elementi sulla base della loro

distanza reciproca, e quindi l'appartenenza o meno a un insieme dipende da quanto l'elemento preso in esame è distante dall'insieme stesso.

Il clustering è ampliamente utilizzato in diverse applicazioni: nell'elaborazione delle immagini, nelle ricerche di marketing e nell'analisi dei dati.

Molti algoritmi sono descritti nella letteratura, come ad esempio il k-means [6], BIRCH [7] che viene impiegato su database di gradi dimensioni ed è in grado di gestire il "rumore" efficacemente, DBSCAN [8] che utilizza una nozione di cluster basata sulla densità che è progettata per scovare cluster di diversa forma, OPTICS [9] che crea un ordinamento potenziato del database che rappresenta la propria struttura clustering basata sulla densità, STING [10] che cattura informazioni statistiche associate a celle nello spazio in modo da riuscire a rispondere alle query e ai problemi di clustering senza ricorrere agli oggetti singoli.

Effettuare un clustering su un'intera traiettoria potrebbe portare però ad un mancato riconoscimento di porzioni di percorso in comune, in quanto nella sua integrità una traiettoria risulta essere lunga e articolata rappresentando così un comportamento d'acquisto che in realtà non è comune a tutti gli individui raggruppati sotto un determinato cluster [5]. Per cui andare a suddividere il cammino in sotto traiettorie può risultare utile nello studiare la condotta delle persone in quei locali in cui si vuole analizzare come i consumatori si relazionano con determinate regioni di interesse.

2. STATO DELL'ARTE

In questo capitolo verranno descritte alcune delle ricerche che sono state effettuate riguardo agli argomenti trattati da questa tesi, in modo da comprendere quali sono state le applicazioni più importanti o più recenti di algoritmi di clustering e quale sono state le ricerche più attuali e innovative riguardanti il campo dello studio del comportamento dei consumatori.

2.1 Clustering

Per quanto riguarda il clustering, un lavoro molto attuale ed interessante è quello esposto nell'articolo [11], che tratta di una ricerca effettuata a Hong Kong, tra il 23 gennaio 2020 e il 28 aprile 2020 in cui si è utilizzato il clustering per suddividere in cluster il numero di persone, pari a 1038, risultate positive al virus Sars-Cov-2 nella regione di Hong Kong durante l'arco temporale sopraindicato.
I casi venivano collegati ai rispettivi cluster basandosi sulla cronologia dei contatti riportati tra le persone, si è arrivati ad avere che un caso positivo risultava collegabile ad almeno uno dei 137 gruppi individuati. La dimensione media di un cluster è 2 e il più grande coinvolgeva 106 casi. Del totale dei contagi 220 erano riconducibili a 22 cluster collegati ad un caso locale, comparati a 89 casi linkabili a 29 cluster innescati da un caso importato. In ogni caso, la maggior parte dei cluster, il 63%, erano classificati come acquisiti all'estero solamente e tra questi per 224 casi non è stato rintracciabile nessun contatto locale per cui l'infezione e il contatto tra loro è stato stabilito fuori dalla regione.
Tra i 505 casi sporadici non collegabili a nessun altro caso, il 90.9% sono riconducibili all'estero, mentre il restante 9.1% sono casi sporadici infettati localmente da quel che si può dedurre basandosi sui loro spostamenti.
Questa ricerca riporta come solo il 31.4% dei casi d'infezione da Sars-cov-2 avvenuti durante il periodo di studio sono avvenuti entro i confini di Hong Kong riconducibili, comunque, a cluster di infezioni provenienti dall'estero o a infezioni sporadiche collegabili a trasmissione locale limitata a contatti nella comunità.

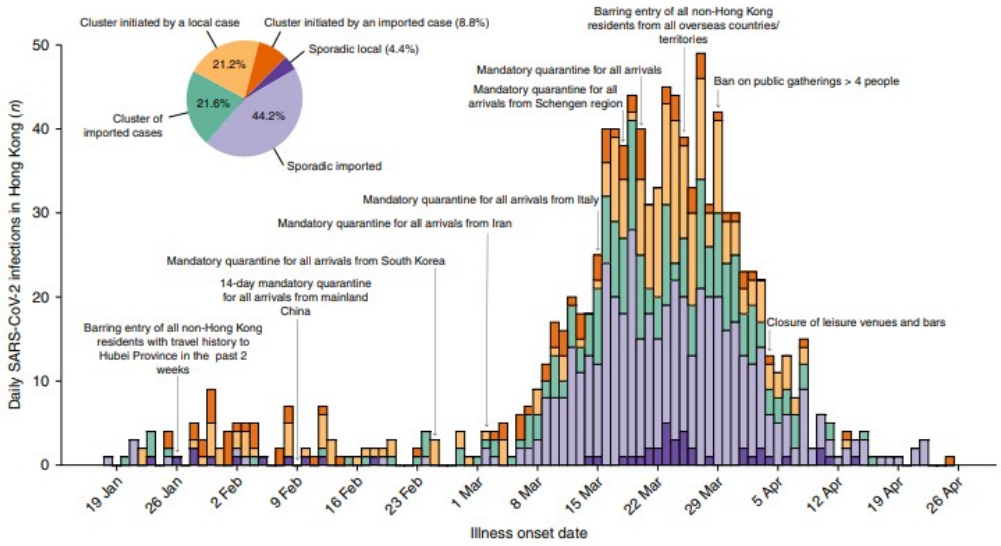

Figura 1

Questo studio mostra un utilizzo interessante e attuale del clustering, grazie al quale si è riusciti ad ottenere una migliore comprensione dei meccanismi di diffusione del virus, che possono essere usati come supporto alle decisioni per quanto riguarda i protocolli da adottare per limitarne la propagazione. Ad Hong Kong, infatti dopo un'impennata nella curva dei contagi dovuta a casi d'importazione è stato vietato l'ingresso nel paese a coloro che non risultavano risiedere lì ed è stato imposto un periodo di quarantena, oltre alle consuete norme di distanziamento sociale.

2.2 Comportamento dei consumatori

Per quanto riguarda l'analisi del comportamento dei consumatori uno studio molto interessante e attuale, poiché è stato effettuato all'inizio del 2020 quando i casi di Covid-19 sono iniziati a crescere anche n Europa, è quello effettuato in Finlandia [12].

Durante la pandemia di Corona virus, comportamenti inusuali, come ad esempio l'acquisto di un numero spropositato di carta igienica, sono stati registrati in tutto il mondo. Questo tipo di condotta ha trovato le sue origini nel momento in cui le persone hanno iniziato a temere un collasso dei mercati e di conseguenza hanno iniziato a comprare grandi quantità di quei beni considerati primari per farne provvista. Questo tipo di condotta è stato indagato sin dall'inizio per studiare il comportamento umano e la sua reazione ad una situazione di stress mondiale come quella che stiamo ancora vivendo al momento.
Basandosi su uno schema S-O-R, ovvero la risposta di un organismo ad uno stimolo, viene proposto un modello strutturale che collega l'esposizione a un numero elevato di informazioni da fonti di diffusione digitali (stimoli ambientali) a due tipi di risposte: l'aumento di acquisti peculiari e la volontà di auto isolarsi. Per testare il

modello proposto sono stati raccolti dati proponendo un questionario online, al quale hanno risposto 211 persone finlandesi, i quali sono stati successivamente analizzati utilizzando PLS-SEM (partial least squares structural equation modeling) [14].

Si è trovato una forte correlazione tra l'intenzione di attuare un distanziamento sociale spontaneo e la tendenza ad effettuare acquisti inusuali, fornendo evidenze empiriche sulla diretta correlazione del comportamento fuori dall'ordinario dei consumatori e la previsione di un periodo di auto isolamento.
I risultati rivelano, inoltre, che un'esposizione a informazioni reperibili online conduce a un sovraccarico di notizie e alla cosiddetta cyberchondria [13], ovvero un comportamento caratterizzato da crescenti preoccupazioni nei confronti di sintomi comuni legati a risultati di ricerche internet su di essi, che può essere innescata anche da una sovra esposizione ad informazioni stessa e da una percezione molto aggravata della situazione che ci circonda.

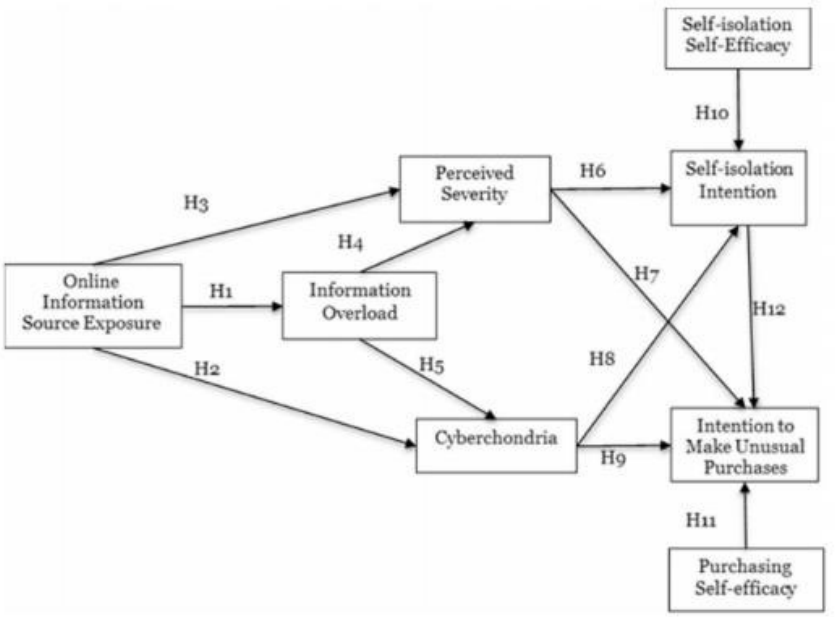

Figura 2

3. MATERIALI E METODI

Questo capitolo si comporrà di diverse parti, in cui verranno descritte le maggiori componenti di questa ricerca.
Come prima cosa si parlerà del dataset su cui si basa questa ricerca, di cosa si compone e dove è stato creato.
In secondo luogo, della tecnologia utilizzata per la raccolta di questi dati e infine verrà esposta la tipologia di clustering scelta e i passaggi di cui si compone l'algoritmo che è stato implementato per l'analisi delle informazioni a nostra disposizione.

3.1 DATASET

Il dataset da noi utilizzato è stato raccolto in un negozio situato in Germania, durante un arco temporale della lunghezza di circa sei mesi, a partire dal 14 gennaio 2016 fino alla fine mese di giugno dello stesso anno.
I dati collezionati sono riferiti all'orario di apertura dell'attività, cioè dalle 07:00 alle 21:59, raccogliendo nel complesso informazioni relative ad un totale di 181339 persone, con una media giornaliera di 1305 individui.
Le informazioni che sono state collezionate comprendono un identificativo del cliente, che viene associato alla persona nel momento in cui entra nel negozio fino ad interrompere la relazione quando l'individuo abbandona il negozio, un identificativo del tag associato a ciascuno dei 250 equipaggiamenti d'acquisto (carrello o cestino della spesa) utilizzati in questa ricerca sperimentale, poi abbiamo dei dati relativi al tempo con la data del rilievo e l'orario con precisione al minuto ed infine abbiamo informazioni relative alla posizione del carrello nello spazio con le coordinate x, y e z con la particolarità di avere z sempre pari a zero poiché si tratta di rilevamenti planari.
Per l'obbiettivo dell'analisi riportata in questa tesi è stata utilizzata solamente una partizione del dataset, relativa a tre giorni lavorativi, i quali per problemi di dimensioni sono stati suddivisi in fasce orarie da un'ora ciascuna.

3.2 TECNOLOGIA ULTRA WIDE BAND

Il sistema di tracking di dati utilizzato per la raccolta delle informazioni usate in questa ricerca è costruito mediante l'uso della tecnologia UWB (ultra-wide band), che è in grado di monitorare le traiettorie dei consumatori nei negozi e inviare i dati collezionati ad un cloud server. Questi dati vengono processati e mantenuti in modo appropriato in un database.
Possono essere utilizzati in modo conveniente per guadagnare informazioni sul comportamento dei consumatori durante l'esperienza d'acquisto.

La tecnologia usata schiera un sistema di localizzazione in tempo reale (RTLS) per accumulare dati relativi alla posizione in tempo reale dei carrelli della spesa e dei cestini.
Dei test attuati in questo negozio nello specifico hanno raggiunto un'accuratezza di 20 cm nella misura della posizione dell'equipaggiamento nella localizzazione in uno spazio chiuso. Le tecnologie wireless convenzionali, ad esempio WLAN, RFID etc, non riescono a ottenere tali valori, quindi questo approccio può essere considerato attuabile nei casi in cui le applicazioni necessitano un alto livello di precisione real time nella localizzazione 2D e 3D.
In aggiunta, il sistema garantisce un'alta autonomia per i tag alimentati da batterie, questo è dovuto da una gestione intelligente dell'energia.
Le informazioni raccolte, ovvero le traiettorie degli acquirenti ma anche per esempio il tempo medio di passeggio e il tempo di attrazione di fronte agli scaffali o ad una determinata categoria di prodotti, possono essere importanti per i commercianti, per migliorare conformemente l'esperienza d'acquisto dei consumatori.
Se è possibile misurare: il tempo di arrivo, direzione d'arrivo e la potenza del segnale.
Questi dati son fondamentali per i rivenditori, in quanto rappresentano il modo in cui i clienti interagiscono con l'ambiente, le aree più visitate del negozio, difficoltà di trovare i prodotti etc. Sono considerate strategie molto efficaci nel campo del marketing, della comunicazione e della progettazione degli ambienti.
In confronto con altre tecnologie, i segnali UWB possono essere trasmessi con una durata più breve, consumando meno energia e operando in un'area più vasta del raggio spettrale. Le applicazioni UWB e RFID possono operare nella stessa zona senza interferenze, ciò dovuto al fatto che i tipi di segnali e raggi spettrali utilizzati sono differenti. Oltretutto, come mostrato nel [16], i segnali UWB sono in grado di passare attraverso muri, dispositivi e vestiti senza perturbazioni.
Il negozio specifico utilizzato per raccogliere i dati immagazzinati nella base di dati sCREEN è un supermercato in Germania, durante le ore di lavoro.
Il dataset contiene informazioni rilevate da sensori installati su carrelli della spesa e cestini di questo specifico supermercato. I dati relativi alla localizzazione generati dall'attrezzatura equipaggiata sono rappresentati come un flusso spaziotemporale di punti, con i parametri (x,y,t). Questi punti vengono poi raggruppati per formare la corrispondente traiettoria, non vengono considerati punti che hanno un tempo d'attrazione minore di 5 secondi di fronte ad uno scaffale. Per capire e processare i dati sulla mobilità è fondamentale costruire la traiettoria. Altre mansioni importanti sono la pulizia dei dati riguardanti la traiettoria, la segmentazione e la compressione, in modo da identificare fermate e movimenti. Per la costruzione della traiettoria di ogni consumatore, nel caso in cui il carrello è rimasto fermo per 5 o più minuti consideriamo che un altro cliente lo abbia preso, quindi abbiamo la creazione di una nuova traiettoria.
Per la raccolta dei dati è stato considerato un periodo di circa 6 mesi. sCREEN è il primo dataset pubblico basato su traiettorie d'acquisto reali raggruppate in circa 6

mesi su 250 carrelli equipaggiati a 1 Hz durante il tempo operativo di un vero supermercato.

3.3 CLUSTERING GERARCHICO

Il clustering gerarchico, nel nostro caso aggregativo, produce raggruppamenti di item (variabili o oggetti) seguendo un processo gerarchico: parte da tutti gli item separati in gruppi individuali [17], e quindi procede iterativamente con l'aggregazione di coppie di gruppi "che si somigliano di più", arrivando al termine a produrre un unico gruppo contenente tutti gli item.
Questa configurazione permette a chi effettua il clustering di determinare qual è lo strato più appropriato per l'applicazione in questione.

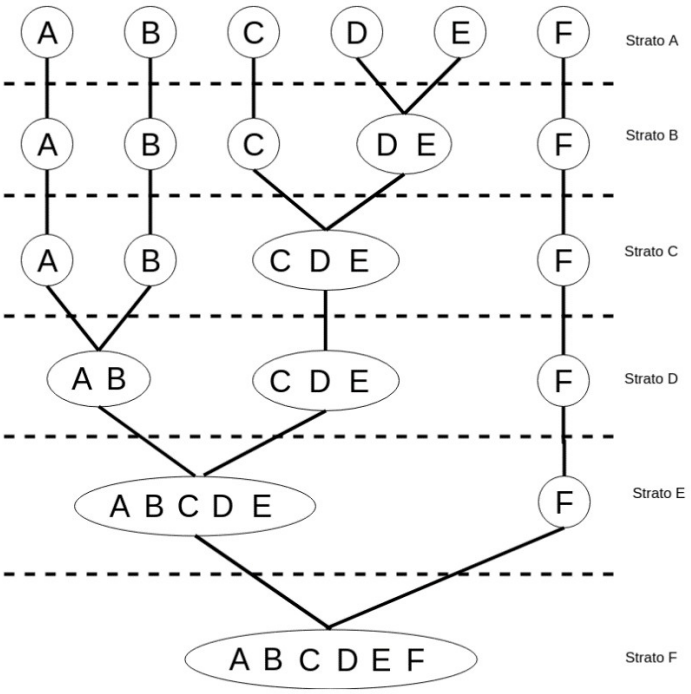

Figura 3

Se P e Q denotano una serie di indici di traiettorie di 2 cluster disgiunti, la distanza media tra i 2 cluster è:

$$d_{avg}(P,Q) = \frac{1}{|P| \cdot |Q|} \sum_{p \in P} \sum_{q \in Q} d(p,q)$$

Nella ricerca di cui si tratta in questo documento il numero di cluster desiderato è stato specificato, piuttosto che scegliere tra i valori ottenibili dall'attuazione dell'algoritmo di clustering sui dati che si avevano a disposizione.
L'intervallo di valori con cui si riesce ad avere come risultato lo stesso numero di cluster è l'indicatore che è alla base della sensibilità della misura delle distanze.

3.4 FILTRAGGIO DATI

I dati da noi considerati, ovvero relativi a tre intere giornate lavorative suddivise in fasce orarie da un'ora ciascuna, prima di essere analizzati mediante l'uso dell'algoritmo di clustering gerarchico sono stati sottoposti ad un filtraggio.
Dalla mole di informazioni a nostra disposizione, tramite l'utilizzo del foglio di calcolo Excel, sono state scremate quelle che vengono definite posizioni fuori limite. Quest'ultime sono quei punti, le quali coordinate ricadono al di fuori della piantina del negozio usata come riferimento da questo esperimento.
Per cui, tramite la funzione filtro di Excel, che permette di filtrare i dati in base al contenuto della casella, sono state eliminate dalla partizione del dataset tutte le righe che possedevano una o entrambe le coordinate x e y negative.

3.5 ALGORITMO

L'algoritmo di clustering gerarchico aggregativo è stato implementato utilizzando l'ambiente di calcolo Matlab [15].
Il codice, mediante il quale è stato realizzato il clustering sulle diverse fasce orarie in cui è stato suddiviso il nostro dataset per essere analizzato è:

```
x=table2array(x);
y= pdist(x);
z=linkage(y);
dendrogram(z)
c=cophenet(z,y)
```

In cui x è la matrice contenente i dati collezionati nel negozio, che tramite la funzione table2array viene trasformata in un array omogeneo.

3.5.1 CALCOLO DELLE DISTANZE

Il primo passaggio di cui si compone l'algoritmo di clustering gerarchico utilizzato per questa ricerca è il calcolo delle distanze.
Si utilizza la funzione *pdist ()* per calcolare la distanza tra ogni coppia di oggetti in un set di dati. Per un set di dati composto da m oggetti, ci sono

m *(m – 1) /2 coppie nel set di dati. Il risultato di questo calcolo è comunemente noto come matrice di distanza. La funzione restituisce queste informazioni sulla distanza in un vettore, Y, dove ogni elemento contiene la distanza tra una coppia di oggetti.
Ci sono diversi metodi che si possono utilizzare per calcolare la distanza, il metodo di default impostato da Matlab è la distanza euclidea, ma se specificato nel codice è possibile scegliere il modo più appropriato per la propria applicazione.
I metodi messi a disposizione da Matlab sono:

Value	Description
`'euclidean'`	Euclidean distance (default).
`'squaredeuclidean'`	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`. Use `DistParameter` to specify another value for `S`.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use `DistParameter` to specify another value for `C`, where the matrix `C` is symmetric and positive definite.
`'cityblock'`	City block distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. Use `DistParameter` to specify a different exponent `P`, where `P` is a positive scalar value of the exponent.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors).
`'correlation'`	One minus the sample correlation between points (treated as sequences of values).
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

Figura 4

3.5.2 CREAZIONE DEI COLLEGAMENTI

Una volta calcolata la vicinanza tra gli oggetti nel set di dati, è possibile determinare come raggruppare gli oggetti nel set di dati in cluster, utilizzando la funzione *linkage ()*.
Quest'ultima prende come input le informazioni sulla distanza generate da *pdist ()* e collega coppie di oggetti vicini tra loro in cluster binari.
La funzione collega, quindi, questi cluster appena formati tra loro e ad altri oggetti per creare cluster più grandi fino a quando tutti gli oggetti nel set di dati originale sono collegati insieme in un albero gerarchico.
Da questa funzione otteniamo in output una matrice, chiamata matrice dei collegamenti di cui le prime due colonne identificano gli oggetti che sono stati collegati mentre la terza colonna contiene la distanza tra questi oggetti.
Linkage utilizza le distanze per determinare l'ordine in cui raggruppa gli oggetti, ovvero item che sono più vicini verranno raggruppati prima degli altri.

Il vettore di distanza Y contiene le distanze tra gli oggetti originali, ma il collegamento deve anche essere in grado di determinare le distanze che coinvolgono i cluster che crea. Per impostazione predefinita, linkage utilizza un metodo noto come collegamento singolo. Tuttavia, sono disponibili diversi metodi:

Method	Description
'average'	Unweighted average distance (UPGMA)
'centroid'	Centroid distance (UPGMC), appropriate for Euclidean distances only
'complete'	Farthest distance
'median'	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only
'single'	Shortest distance
'ward'	Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only
'weighted'	Weighted average distance (WPGMA)

Figura 5

3.5.3 DENDOGRAMMI

Un metodo per comprendere meglio come viene generato l'albero gerarchico è visualizzarlo.
Matlab consente di graficare il dendogramma utilizzando la funzione *dendrogram()*, la quale prende in ingresso la matrice dei collegamenti e restituisce un'immagine raffigurante la struttura dei cluster.

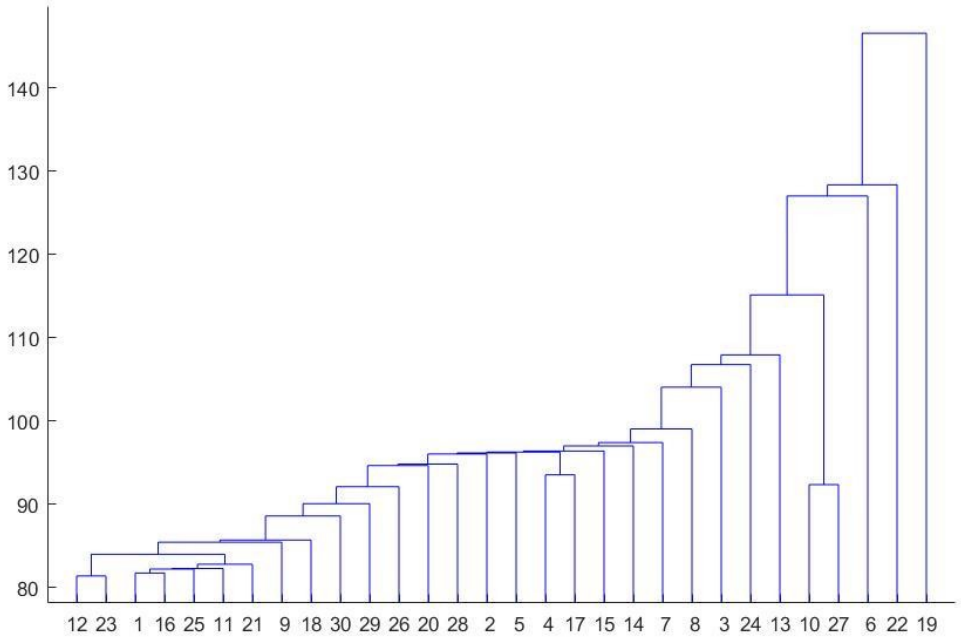

Figura 6

3.5.4 VERIFICA DELLA DISSIMILARITA'

Dopo aver ottenuto l'albero gerarchico e averlo graficato è necessario comprendere se i collegamenti generati siano consistenti.
Per fare ciò bisogna ricordarci che l'altezza del collegamento rappresenta la distanza tra i due cluster contenenti i due oggetti separati, nominata nella letteratura anche come *distanza cophenetica*.
Quindi un metodo per valutare la qualità dei collegamenti generati dalla funzione linkage () è quello di confrontare la distanza cophenetica con la reale distanza presente tra i due oggetti precedentemente calcolata con la funzione *pdist* ().
Se il clustering effettuato è consistente, i collegamenti tra gli oggetti mostrano uno forte correlazione con le distanze tra gli item del vettore di distanza.
La funzione *cophenet* () confronta questi due insiemi di valori e calcola la loro correlazione, restituendo un valore chiamato *coefficiente di correlazione cophenetica*, più il valore è prossimo all'uno più la soluzione di clustering che è stata implementata riflette al meglio il contenuto del dataset analizzato.

4. RISULTATI E DISCUSSIONI

In questo capitolo della tesi si tratterà dei risultati ottenuti dall'applicazione dell'algoritmo di clustering gerarchico sui dati a disposizione, in particolar modo si parlerà della qualità ottenuta implementando il codice esposto nel capitolo precedente, facendo uso della funzione cophenet () e dell'analisi che ne è scaturita successivamente ai valori ottenuti.
Si analizzerà come la grandezza delle matrici utilizzate influenzi la qualità del clustering ottenuto e in che modo diversi metodi di calcolo delle distanze possano portare ad avere alberi gerarchici che rappresentino meglio le informazioni contenute dal dataset raccolto durante l'esperimento.

4.1 ANALISI DEI COEFFICIENTI DI CORRELAZIONE

Successivamente alla creazione dell'albero gerarchico per ciascuna fascia oraria considerata per la realizzazione di questa ricerca, si è andato a calcolare il coefficiente di correlazione cophenetica col fine di valutare la qualità dei clustering ottenuti in ogni periodo indagato.
Quindi applicando la funzione cophenet (), dando in ingresso la matrice della distanza e la matrice dei collegamenti si è computato il valore del coefficiente che per comodità nelle seguenti pagine verrà nominato c.

Ricordando che il valore di c è tanto migliore tanto più esso è prossimo all'uno, abbiamo ottenuto che sulle 45 fasce orarie da un'ora ciascuna, in cui la partizione del dataset è stata suddivisa, 31 hanno riportato un valore cophenetico maggiore di 0.6, quindi in quelle fasce orarie il clustering ottenuto è stato considerato soddisfacente, e non si è indagato ulteriormente la qualità dei collegamenti in base al metodo di calcolo delle distanze.
Delle restanti 14 fasce orarie che sono state ricavate dalla suddivisione, invece procederemo con un'analisi più dettagliata dei metodi di calcolo della distanza e dei rispettivi coefficienti di correlazione cophenetici ottenuti, in quanto il valore della loro c, ricavato utilizzando la distanza di default di Matlab ovvero la distanza euclidea, è inferiore a 0.6.
Per cui si è effettuata un'analisi su questi dati che consisteva nel calcolare le distanze con i diversi metodi messi a disposizione da Matlab e andare iterativamente poi a rigenerare dei nuovi collegamenti, in modo da verificare se con la modifica della distanza si riuscisse ad ottenere un clustering che rappresentasse meglio i dati da noi utilizzati.
Considerando le 3 giornate considerate abbiamo che 5 periodi sono riferiti alla prima giornate, altri 5 sono di riferimento alla seconda giornata mentre solamente 4 sono relative alla terza giornata considerata.

Nel dettaglio:

30/06: si tratta delle fasce orarie ➔ 13-11-10-09-08
29/06: si tratta delle fasce orarie➔ 16-15-14-13-08
28/06: si tratta delle fasce orarie➔ 14-13-12-10

4.1.1 30/06

Nel corso di questa giornata sono stati individuati 5 orari i quali coefficienti di correlazione cophenetica sono da considerarsi non soddisfacenti.
La prima fascia oraria ad essere risultata problematica è quella delle ore 08,

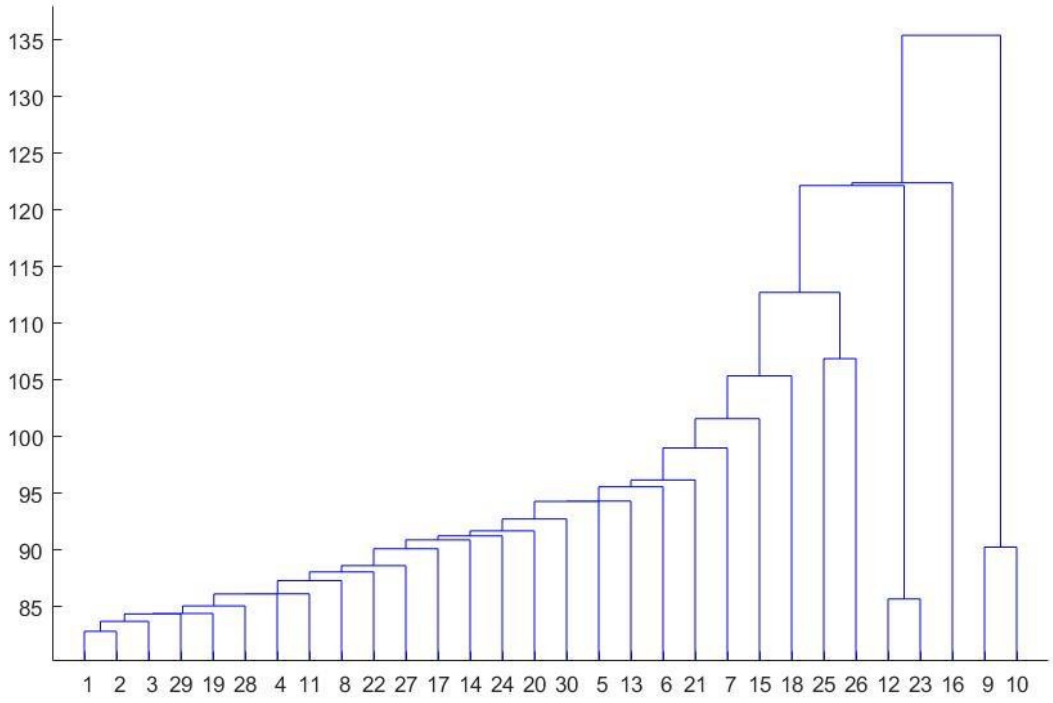

Figura 7

In cui con l'utilizzo del metodo standard di calcolo della distanza si otteneva un coefficiente di correlazione cophenetica pari a c=0.4916. Dovendo essere c il più vicino possibile ad uno, in questo caso si aveva un valore inferiore al 50%, per cui si è andato nuovamente a calcolare la distanza utilizzando la funzione pdist(x,''), dove tra gli apici si è andato a specificare il metodo scelto per il computo.

Nel caso di questa fascia oraria si sono notati miglioramenti in due casi, rispettivamente con il metodo standard euclideo e con il metodo Cityblock:

Euclideo standard (seuclidean): c=0.4918
Cityblock: c=0.4959

Come possiamo notare abbiamo che rispetto al metodo di default il numero è stato incrementato, ma la crescita è stata minima ed inoltre il valore ottenuto nel miglior caso risulta essere ancora inferiore a 0.5.

La seconda fascia oraria in data 30/06 in cui abbiamo riscontrato un valore cophenetico inferiore allo 0.6 è il periodo delle ore 09,

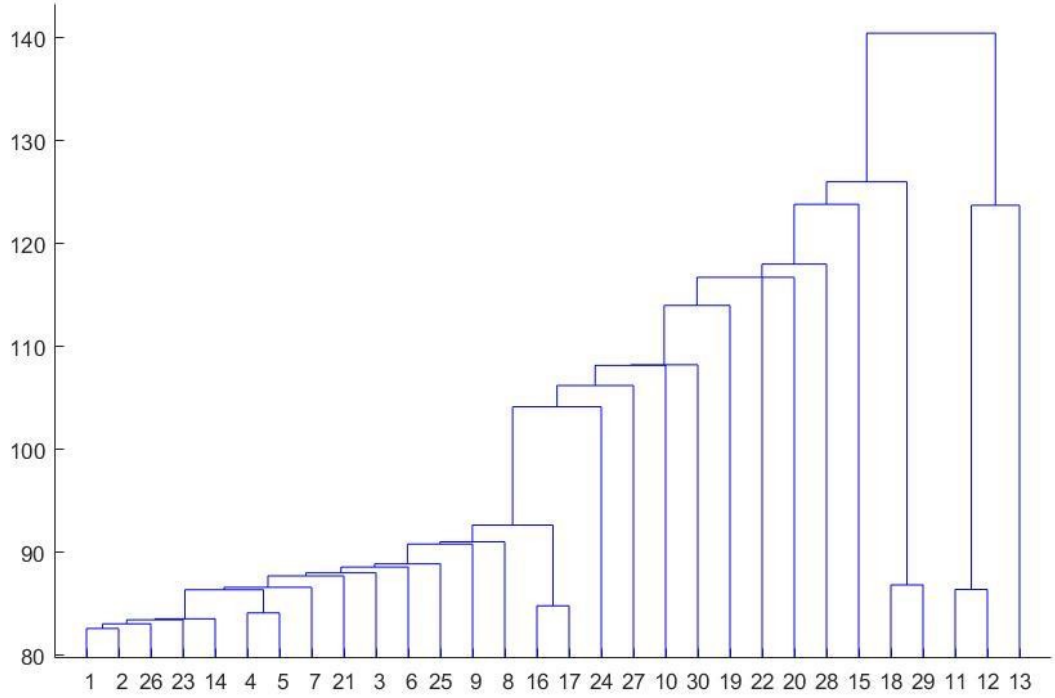

Figura 8

Anche su questo clustering riscontriamo una qualità dell'albero gerarchico inferiore allo 0.5, infatti dall'implementazione della funzione cophenet() sui dati delle 09 si ottiene una c=0.4355.
Applicando la stessa prassi anche in questo caso si sono ottenuti miglioramenti mediante l'utilizzo del metodo Mahalanobis, Chebychev e Cosine, con i rispettivi risultati:

Mahalanobis: c=0.4638

Chebychev: c=0.4696
Cosine: c=0.4959

Da notarsi un lieve miglioramento nel valore del coefficiente di correlazione cophenetico che comunque rimane al di sotto dello 0.5 .

Il successivo arco temporale, relativo alla giornata lavorativa del 30/06 che andremo a considerare in quanto con l'utilizzo della distanza euclidea riporta un valore di c minore al 60%, è quello che va dalle 11.00 alle 11.59 in cui il valore di c=0.5530.

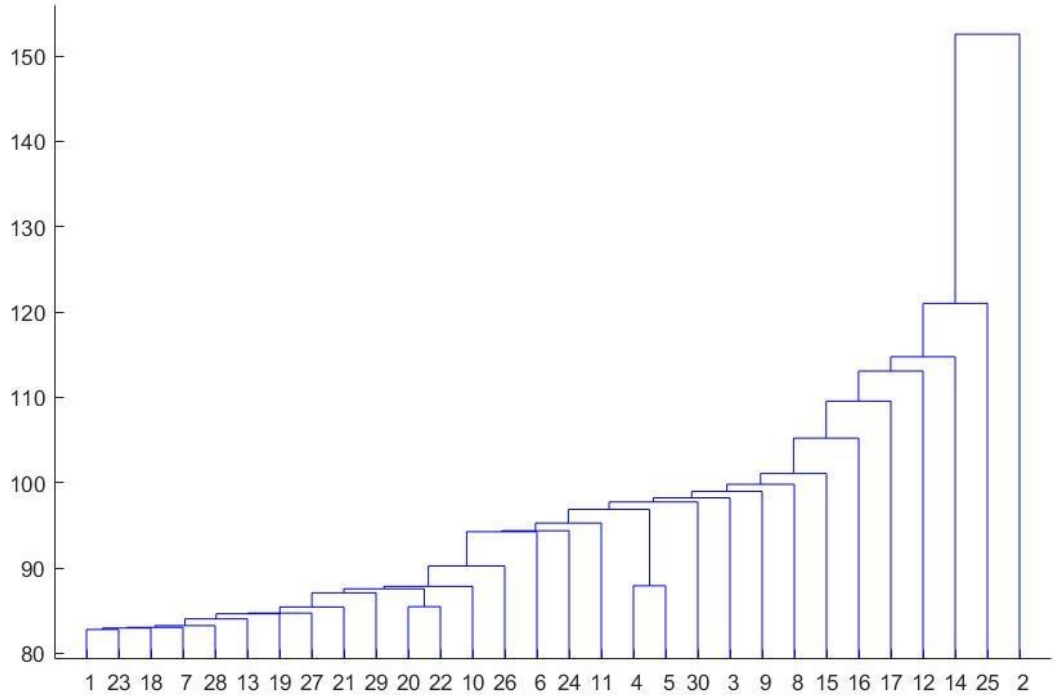

Figura 9

In questo caso il valore del coefficiente è maggiore di 0.5 quindi risulta essere migliore delle due fasce orarie analizzate precedentemente, ma si è analizzato comunque il risultato del clustering gerarchico effettuato con l'utilizzo di diverse distanze.
I metodi usati in questo caso che hanno permesso di vedere un miglioramento nel valore del coefficiente di correlazione cophenetica sono stati il metodo standard euclideo, il metodo di Chebychev e il metodo di Cosine, con i rispettivi risultati:

Standard euclideo (seuclidean): c=0.5552
Chebychev: c=0.5748

Cosine: c=0.6275

Il valore di c, in questo caso raggiunge un valore superiore a 0.6 con l'utilizzo della distanza di Cosine.
Quindi, se considerassimo questo metodo nell'implementazione della pdist(x,'cosine') otterremmo un clustering soddisfacente per questa fascia oraria nello specifico.

L'ultima fascia oraria, valutata in questa giornata è quella che va dalle 13.00 alle 13.59.
In questa fascia otteniamo dall'implementazione dell'algoritmo, nel caso di distanza euclidea di default, un coefficiente di correlazione cophenetica del valore di c=0.5524.

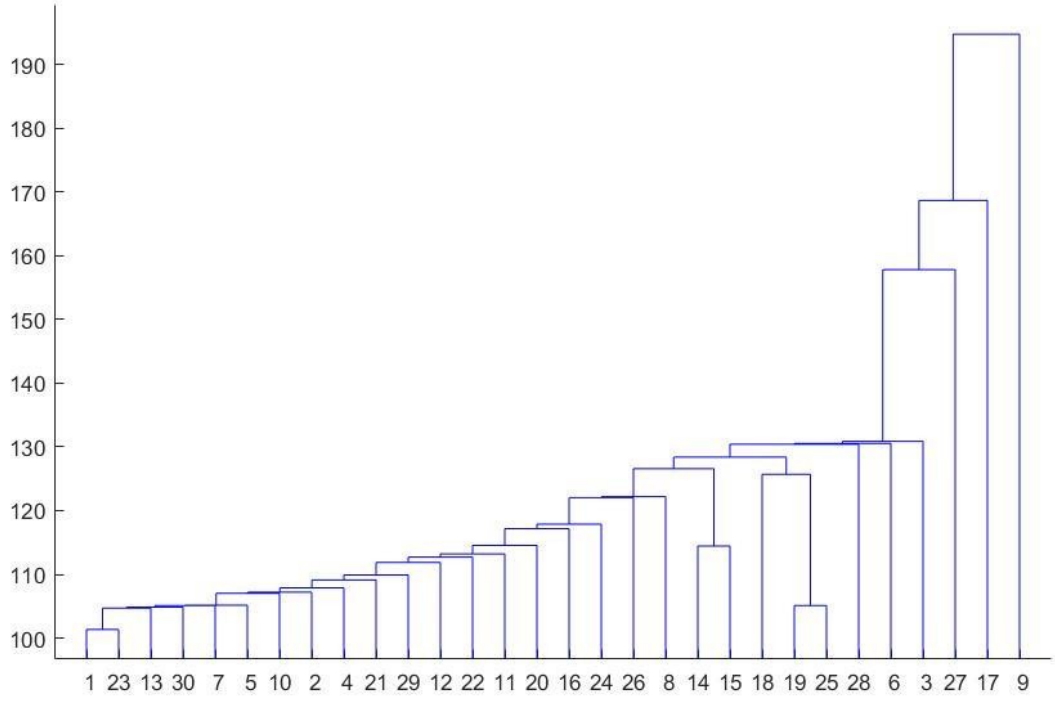

Figura 10

In questo caso otteniamo che il coefficiente ha un valore compreso tra 0.5 e 0.6, che è migliore rispetto ad altre fasce orarie già considerate, ma sono state comunque indagate per valutare un eventuale miglioramento mediante l'utilizzo di altre tipologie di calcolo delle distanze.
Le metodologie che per questa fascia oraria hanno apportato un incremento al valore cophenetico sono il metodo euclideo standard, il metodo di Mahalanobis e il metodo di Chebychev, con i seguenti risultati:

Euclideo standard(seuclidean): c=0.5658
Mahalanobis: c=0.5671
Chebychev: c=0.5721

In questo caso il miglior albero gerarchico lo otteniamo utilizzando la distanza di Chebychev, ma l'incremento ottenuto a differenza della fascia oraria precedentemente analizzata non è significativo. Si mantiene, infatti, un valore di c compreso tra 0.5 e 0.6.

4.1.2 29/06

Durante l'orario lavorativo preso in considerazione nella giornata lavorativa del 29/06 sono state individuate 5 fasce orarie in cui il coefficiente di correlazione cophenetica è inferiore a 0.6.

La prima fascia interessata da questa analisi è stata quella delle 08, in cui il coefficiente è pari a c= 0.5778.

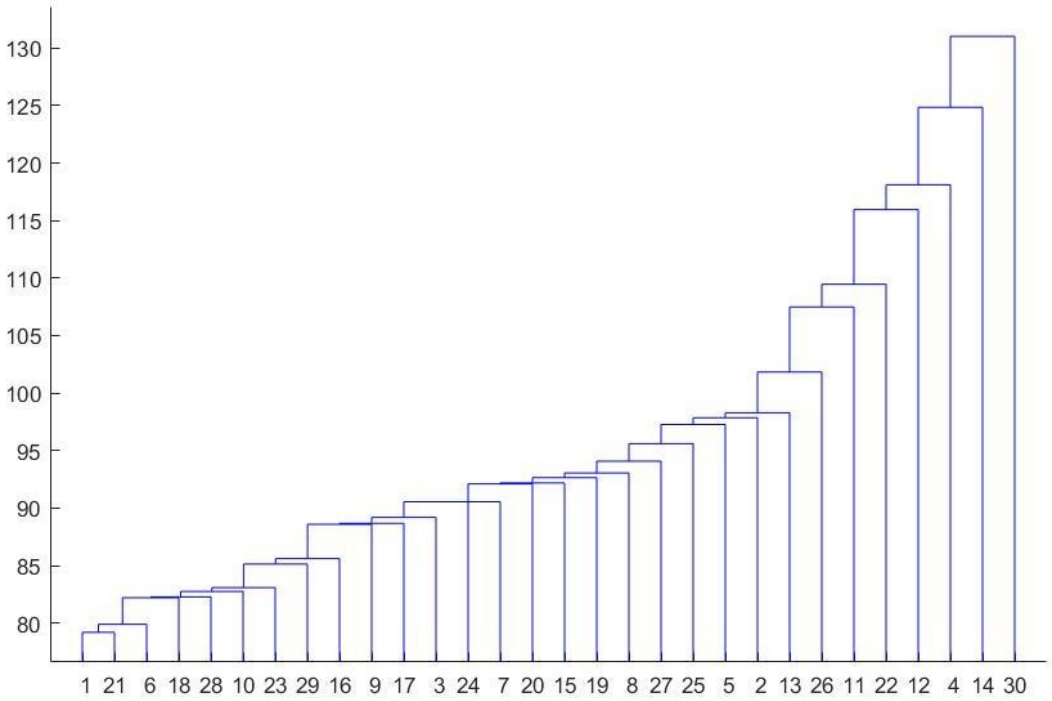

Figura 11

In questo caso la funzione linkage () dà in output un albero gerarchico che descrive quasi in modo soddisfacente i dati riferiti a questo arco temporale, infatti il valore della c risulta essere molto prossimo a 0.6.
Per vedere se con un metodo delle distanze differente si riusciva ad ottenere un valore del coefficiente cophenetico superiore a 0.6 è stata

applicata la procedura analoga alle altre fasce orarie ricalcolando la matrice delle distanze con metodi diversi. In questo caso si sono ottenuti risultati maggiori alla c, calcolata con il metodo di default di Matlab, solamente con l'utilizzo del metodo di Mahalanobis, il quale restituisce un valore pari a:

Mahalanobis: c=0.6042

Per questa fascia oraria si ottiene un miglioramento minimo che ci permette di portare il coefficiente leggermente sopra la soglia dello 0.6.

La seconda fascia oraria presa in analisi per valutare il valore del coefficiente di correlazione cophenetica è quella che va dalle 13.00 alle ore 13.59 della giornata del 29/06.
Il valore di c, calcolato con il metodo euclideo, è pari a 0.5419.

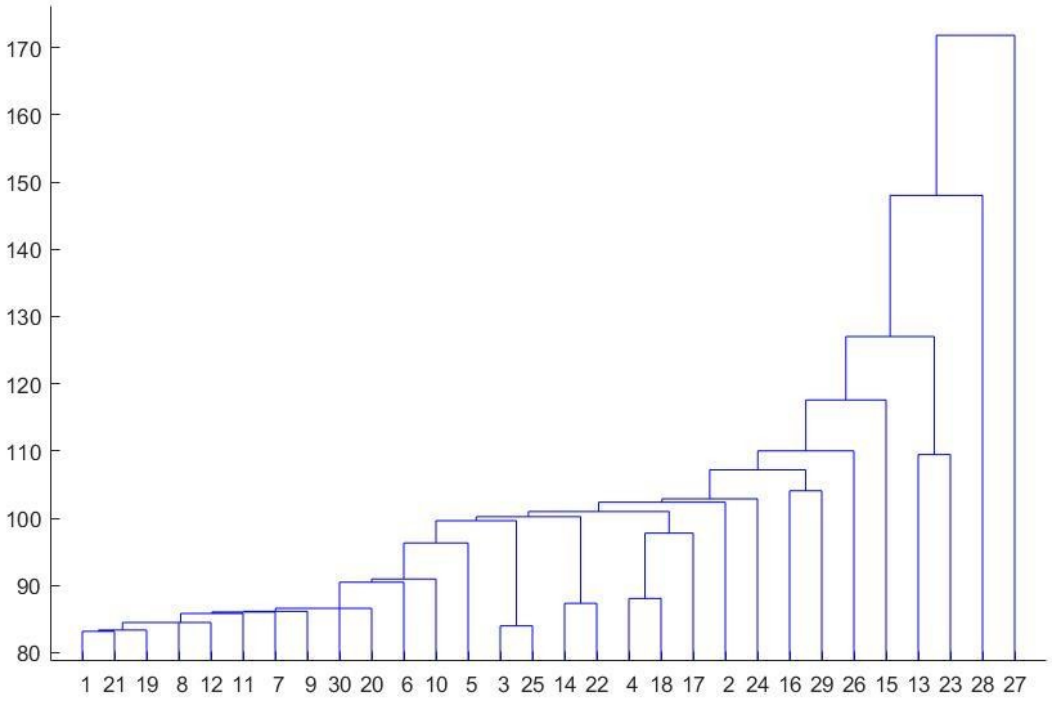

Figura 12

Questo albero gerarchico è riferito al clustering mediante le distanze euclidee, mentre i risultati ottenuti dall'applicazione iterativa dello stesso algoritmo ma con i diversi metodi di calcolo della distanza sono:

Cityblock: c=0.5491

Chebychev: c=0.5739

In questo caso il miglioramento ottenuto con l'utilizzo di diversi metodi di calcolo non ci garantisce un risultato maggiore di 0.6, infatti il miglior valore che andiamo ad ottenere lo si ha usando la distanza di Chebychev, ma anche se si parla di un incremento, non è di fatti un incremento significativo in quanto fa aumentare il valore del coefficiente di correlazione solo di 0.03.

L'arco temporale successivo che andremo ad analizzare è quello che va dalle 14.00 alle 14.59.
Nel clustering dei dati relativi a quest'ora si riscontra un coefficiente di correlazione cophenetica pari a c=0.5628.

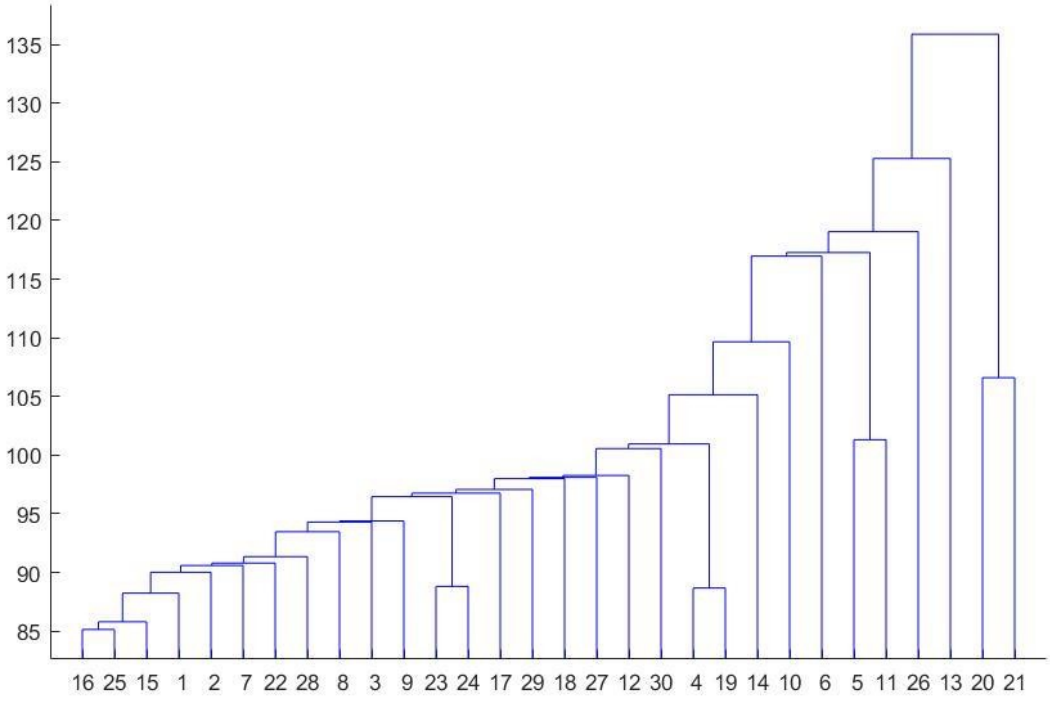

Figura 13

Questo albero gerarchico rappresenta dei collegamenti sulla base di un calcolo delle distanze effettuato con il metodo euclideo, che come sopra indicato, risulta avere c minore di 0.6, per cui dopo aver ricalcolato le distanze mediante l'utilizzo degli altri metodi a disposizione su Matlab, si è giunti alla conclusione che solamente con la distanza di Mahalanobis si otteneva un valore maggiore di quello ottenuto di default, mentre con gli altre opzioni di calcolo non si aveva altro che peggioramenti del coefficiente, quindi abbiamo che:

Mahalanobis: c=0.5647

Come si nota, l'incremento è infinitesimo per cui non è significativo ed è indifferente l'utilizzo di questo metodo o quello euclideo di default.

Il periodo seguente che andremo ad analizzare è l'arco temporale che va dalle 15.00 alle 15.59, in cui si ottiene un coefficiente di correlazione cophenetica pari a c=0.5385.

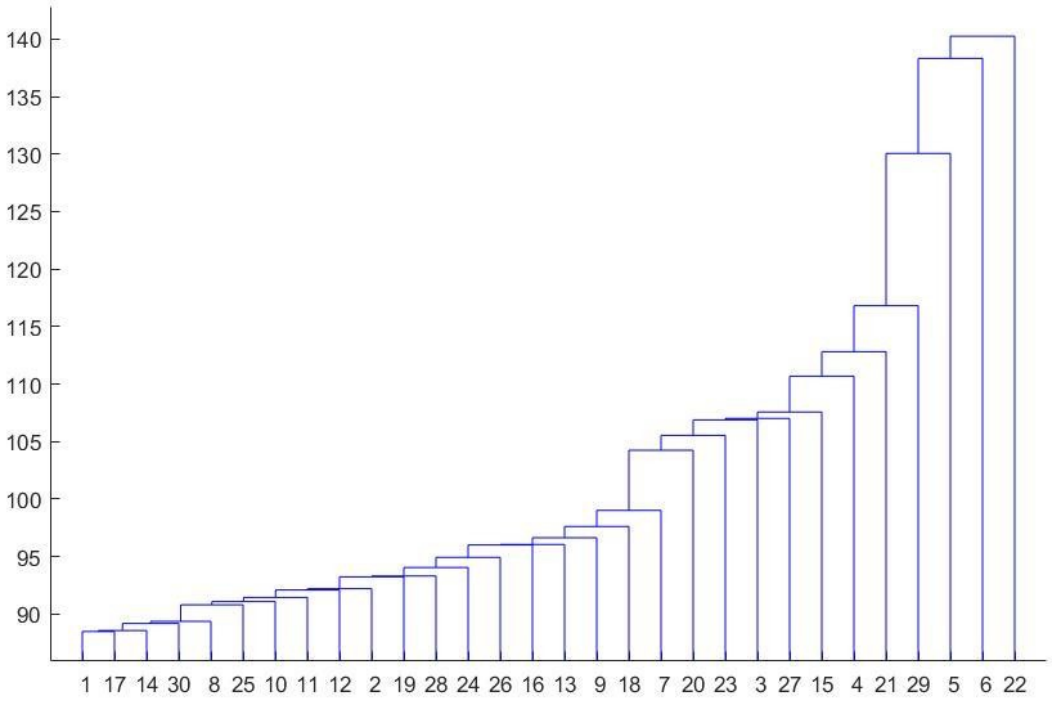

Figura 14

Successivamente all'analisi effettuata sui dati orari mediante il calcolo delle distanze e la successiva rigenerazione degli alberi gerarchici con la funzione linkage (), si denotano valori della variabile c incrementati con il metodo euclideo standard, con il metodo di Mahalanobis, quello di Chebychev ed infine quello di Cosine. Successivamente si riportano i valori:

Euclideo standard(seuclidean): c=0.5387
Mahalanobis: c=0.5396
Chebychev: c=0.5628
Cosine: c=0.6093

Dall'uso delle diverse distanze notiamo che su questa fascia oraria l'utilizzo della specifica Cosine nella funzione pdist (), genera un miglioramento nella qualità del clustering effettuato, facendo aumentare il valore del coefficiente cophenetico fino a 0.6093.

L'ultima fascia oraria relativa alla giornata del 29/06 che andremo a considerare è quella delle ore 16.
In questa fascia oraria l'albero gerarchico risultante dalla funzione linkage() con distanza euclidea di default, ha un coefficiente cophenetico pari a c=0.5293 .

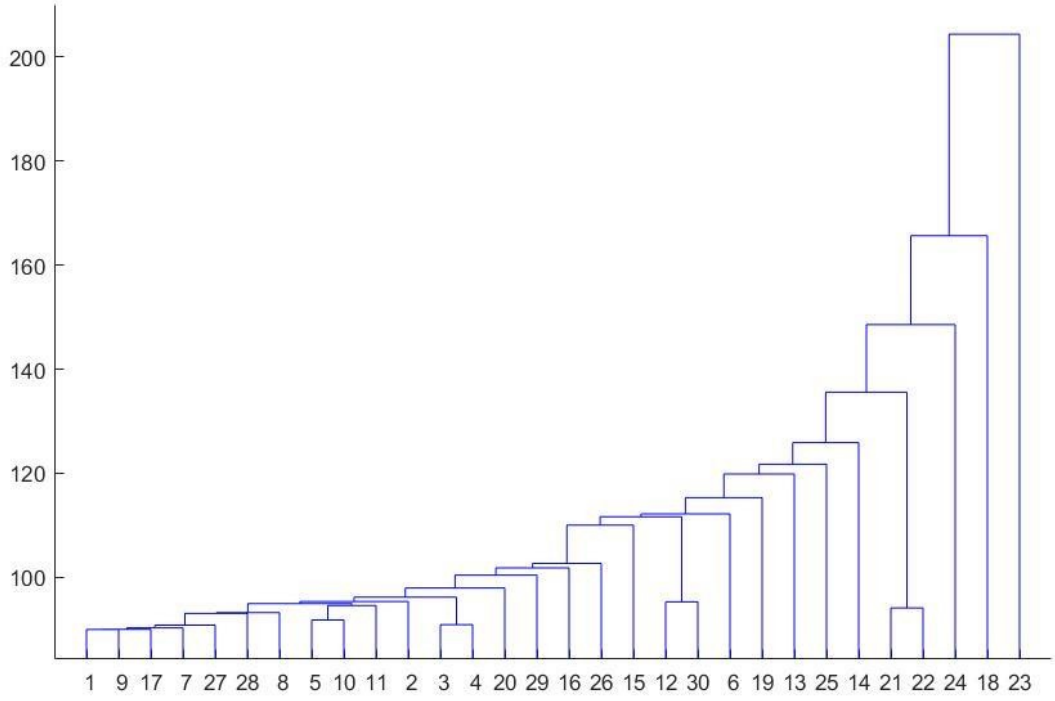

Figura 15

In questo caso, il valore di c non risulta migliorare molto, infatti delle distanze a disposizione solamente due di queste apportano un incremento alla qualità del dendrogramma, ovvero il metodo euclideo standard e il metodo di Cosine, di seguito vengono riportati i risultati prodotti:

Euclideo standard (seuclidean): c=0.5423
Cosine: c=0.5530

Come si può notare, anche nel miglior caso ottenuto il valore del coefficiente rimane compreso tra 0.5 e 0.6, ma comunque è maggiore del 50%, a differenza di alcuni casi precedentemente trovati nell'indagine.

4.1.3 28/06

Nella giornata relativa al 28/06, sono sati individuati 4 periodi in cui il coefficiente di correlazione cophenetica risulta inferiore a 0.6 e quindi sono stati successivamente analizzati ambiando i metodi di calcolo delle distanze.

La prima fascia oraria interessata, di questa giornata, è quella che va dalle ore 10.00 alle 10.59, in questo caso otteniamo con l'uso della distanza euclidea un valore c=0.5194.

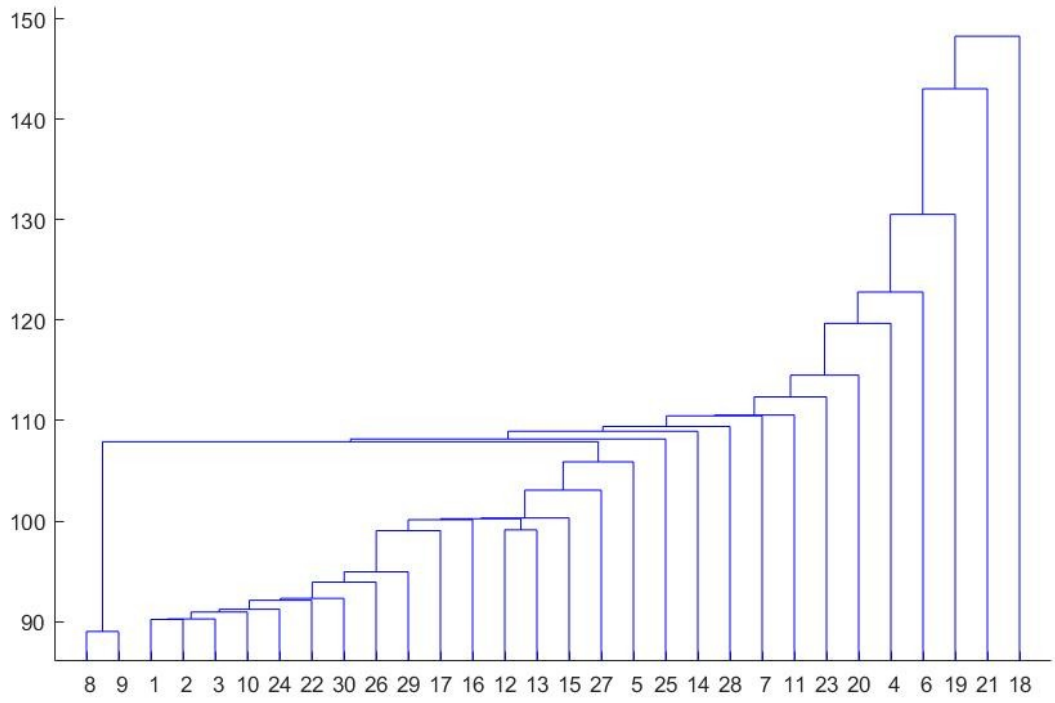

Figura 16

Abbiamo quindi un valore superiore a 0.5, ma non di molto per cui si procede ricalcolando le distanze con la funzione pdist () e otteniamo che solo in un caso

abbiamo un accrescimento del valore cophenetico, ovvero utilizzando la distanza di Mahalanobis:

Mahalanobis: c=0.5658

L'arco temporale successivamente considerato a causa del proprio valore c inferiore a 0.6 è quello che va dalle 12.00 alle 12.59.
Durante questo periodo il clustering effettuato riporta un coefficiente di correlazione cophenetica pari a c=0.5526.

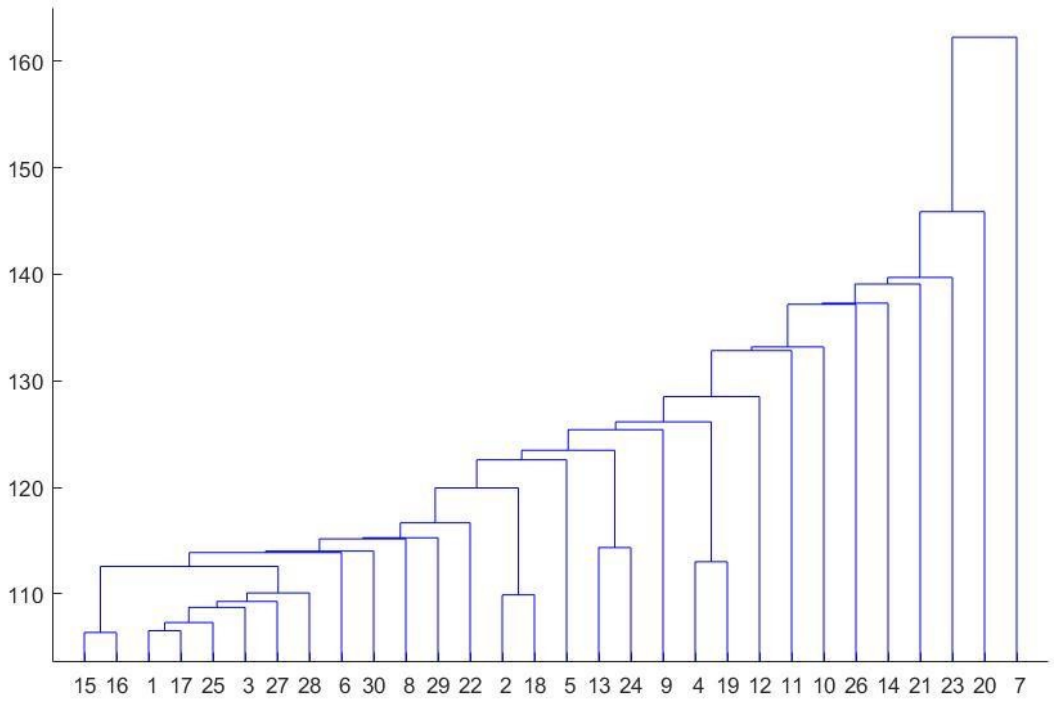

Figura 17

Otteniamo quindi un coefficiente c compreso tra 0.5 e 0.6, quindi si è andato a ricalcolare le distanze e successivamente i collegamenti in modo da osservare se ci fosse un metodo con cui si ottenesse un risultato più vicino al valore ideale, pari a uno.
In questo caso i metodi che apportano un miglioramento sono quello euclideo standard e il metodo Cityblock:

Euclideo standard (seuclidean): c=0.5544
Cityblock: c=0.5684

Neanche in questo caso otteniamo un risultato maggiore di 0.6, ma comunque abbiamo un valore superiore alla metà del valore massimo.

La fascia oraria che andremo a considerare adesso è quella riguardante il periodo che va dalle ore 13.00, alle ore 13.59.
In questa fascia oraria la funzione cophenet (), dà in output un valore di coefficiente cophenetico pari a c=0.5399.

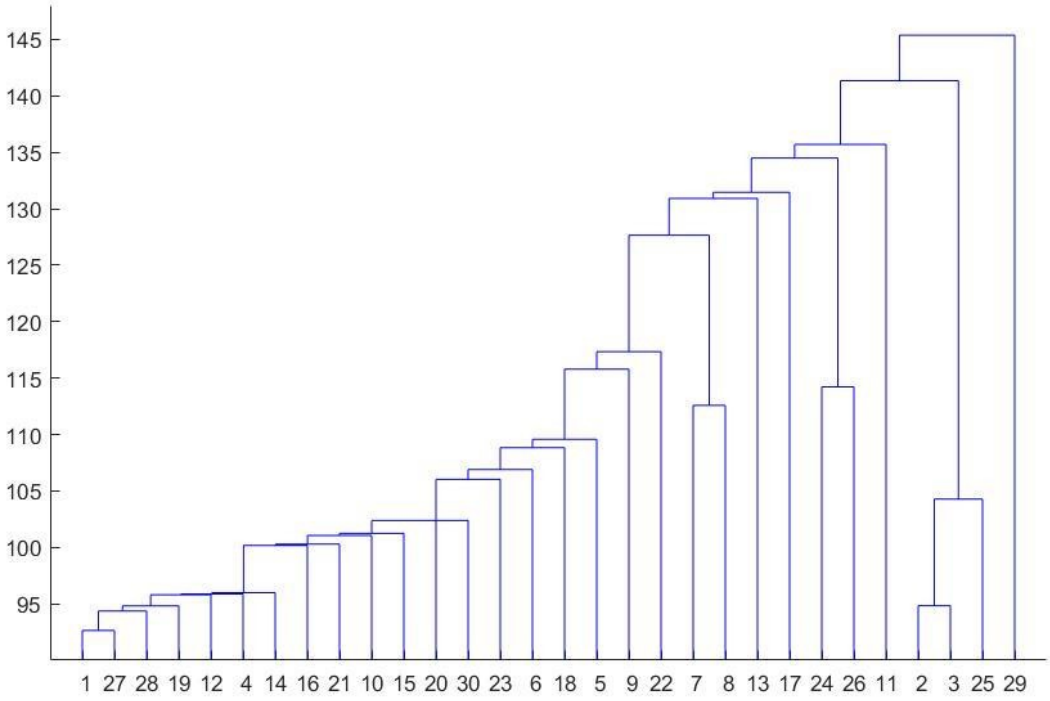

Figura 18

Dopo aver ottenuto tali dati, si è ulteriormente indagato questo periodo per controllare se con distanze differenti si riuscisse ad ottenere una qualità del clustering gerarchico migliore, i metodi con cui ciò si verifica sono le distanze di Mahalanobis quella di Cosine, successivamente si riportano i risultati ottenuti:

Mahalanobis: c=0.5630
Cosine: c=0.6880

In questo periodo abbiamo che il risultato migliore si ha con l'utilizzo del metodo della distanza di Cosine, che ci permette di ottenere un valore superiore a 0.6, quasi prossimo a 0.7 quindi in questa occasione a differenza della maggior parte dei risultati ottenuti in questa analisi abbiamo un risultato che subisce un incremento significativo.

L’ultima fascia oraria che andiamo a considerare in questa terza giornata, nonché l’ultima fascia oraria impiegata in questa analisi oraria è quella compresa tra le ore 14.00 e le ore 14.59.
In questo arco temporale otteniamo un valore del coefficiente di correlazione cophenetica inferiore a 0.5, nel dettaglio il valore della c è pari a c=0.4744.

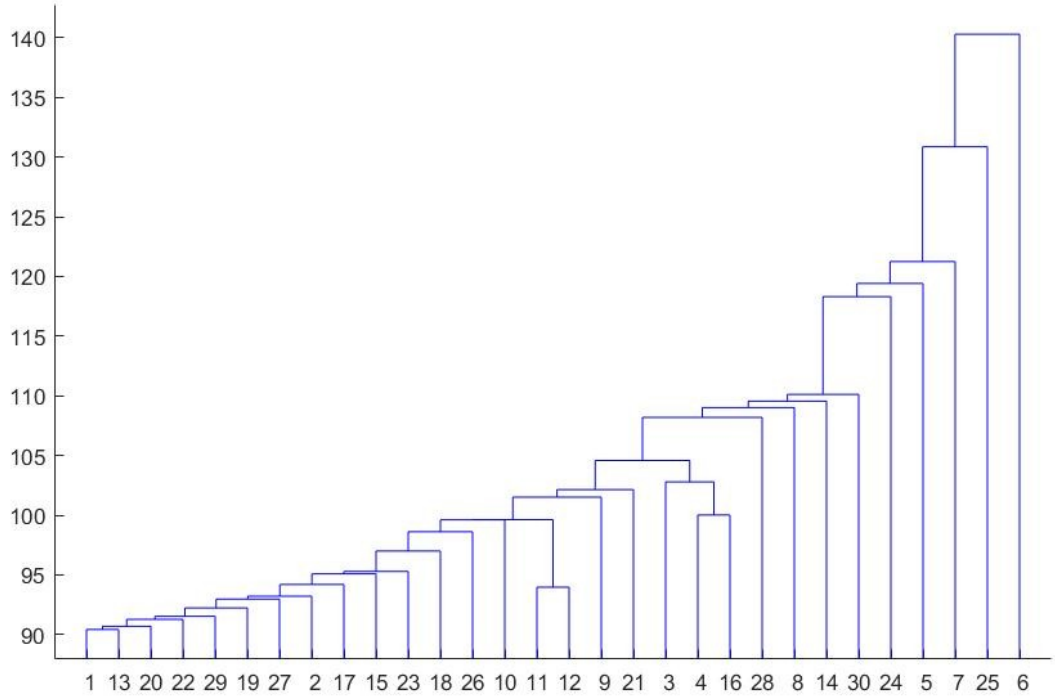

Figura 19

In questo caso il valore ottenuto è inferiore a 0.5, per cui è stata necessario un ricalcolo delle distanze e una successiva rigenerazione degli alberi gerarchici, i risultati ottenuti dalla riapplicazione dell’algoritmo sono incrementati solo nei casi in cui si usassero il metodo eulideo standard oppure il metodo di Cosine, i valori ottenuti sono:

Euclideo standard (seuclidean): c=0.4846
Cosine: c=0.4933

Il valore del coefficiente di correlazione cophenetica, in questo arco temporale, subisce un lieve incremento, ma oltre ad essere molto basso in termini numerici non ci restituisce un albero gerarchico soddisfacente, infatti il valore rimane al di sotto di 0.5, in tutti i casi osservati.

4.2 ANALISI DEI DATI A 30 MINUTI

Dopo aver eseguito l'analisi dei coefficienti di correlazione sulle fasce orarie, si è notato che buona parte di questi valori risulta essere superiore a 0.5, quindi comunque maggiore della metà del valore massimo che c può assumere.
Per cui in questa sezione andremo ad analizzare quelle fasce orarie in cui, nonostante fosse stato applicato un metodo di calcolo della distanza diverso non si è riusciti ad ottenere un albero gerarchico qualitativamente soddisfacente, ovvero quegli archi temporali in cui i coefficienti risultano ancora inferiori a 0.5, anche dopo l'analisi delle distanze.
I periodi temporali a cui ci si riferisce sono relativi a soli due giorni analizzati, nel dettaglio si tratta delle ore 08 e 09 del giorno 30/06 e delle ore 14 del 28/06.
In questi periodi vengono riscontrati valori compresi tra lo 0.4 e lo 0.5 per cui sono stati analizzati singolarmente.
Nell'analizzarli si è notata una cosa molto importante, che si suppone sia la ragione per cui questi periodi avevano una qualità del clustering inferiore agli altri; ovvero osservando le matrici della distanza ottenute utilizzando il metodo euclideo di default si nota che queste fasce orarie sono quelle caratterizzate da matrici dimensionalmente più grandi, rispettivamente:

- La matrice delle ore 08 del 30/06 ha dimensione 2.3 Gb
- La matrice delle ore 09 del 30/06 ha dimensione 1.6 Gb
- La matrice delle ore 14 del 28/06 ha dimensione 1.3 Gb

Davanti a questi risultati si è ipotizzato che le fasce orarie in cui si verificavano problemi sulla qualità siano le ore in cui il negozio abbia avuto la maggior affluenza di persone, per cui si è deciso di suddividere l'arco temporale da un ora in 2 sottoinsiemi da 30 minuti ciascuno, per sperimentare se con un numero minore di dati si riuscisse ad ottenere un raggruppamento che riflettesse in modo più preciso le informazioni deducibili dai dati relativi a queste fasce orarie del dataset.

Quindi nelle pagine successive osserveremo in che modo si comporta il clustering su fasce orarie da 30 minuti, che sono riportate nelle righe a seguire:

08.00→ 08.30

08.31→ 08.59

09.00→ 09.30

09.31→ 09.59

Per quanto riguarda la giornata del 30/06, invece per il 28/06 abbiamo

14.00→ 14.30

14.31➔ 14.59

4.2.1 ORE 08

Consideriamo in primis la fascia oraria che va dalle 08.00 alle 08.30.
In questo orizzonte temporale abbiamo che il coefficiente di correlazione cophenetico, calcolato su una matrice della distanza euclidea, ha valore pari a c=0.5347.

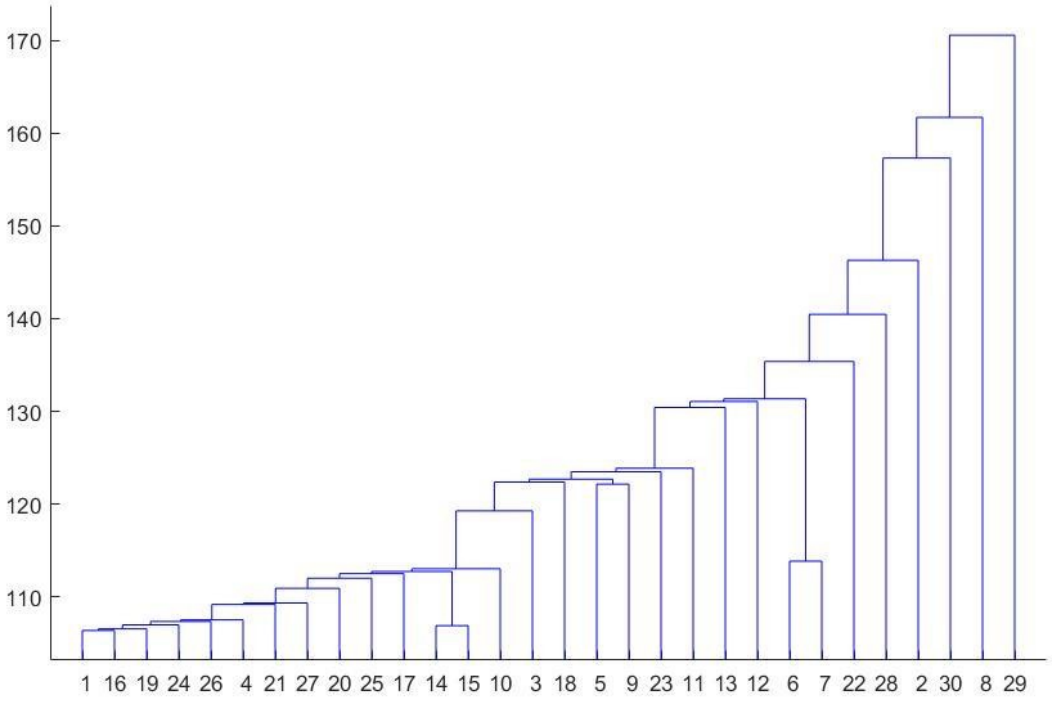

Figura 20

Dato il risultato compreso tra 0.5 e 0.6 del coefficiente, si è effettuata un ulteriore analisi sulle distanze, come precedentemente fatto sui periodi da un'ora.

Ricalcolando le distanze utilizzando i diversi metodi a disposizione si sono ottenuti valori cophenetici migliori con l'utilizzo dei seguenti metodi:

Euclideo standard (seuclidean): c= 0.5377

Mahalanobis: c=0.5562

Chebychev: c=0.5747

Cosine: c=0.6160

In questo caso il miglior valore si ha con il metodo di Cosine:

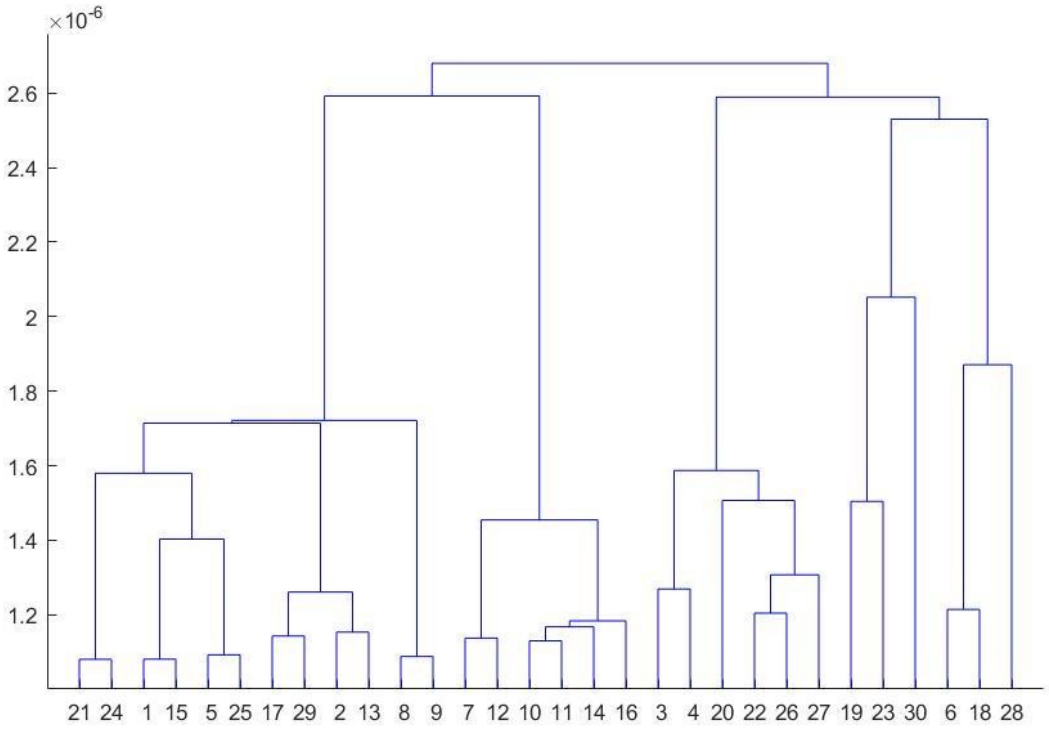

Figura 21

Infatti, si riesce ad ottenere un coefficiente cophenetico che supera lo 0.6 e quindi si ottiene una rappresentazione migliore della prima parte dell'ora scindendo in 2 insiemi la fascia oraria.

Procediamo ora con la medesima analisi effettuata però sulla seconda partizione delle ore 08, ovvero 08.31➔ 08.59.
Il coefficiente ottenuto dall'implementazione della funzione cophenet () con la matrice della distanza computata con il metodo di default di Matlab, equivale a c=0.5062.

In questo caso si ottiene un risultato che è migliore rispetto a quello avuto nell'analisi del lasso temporale nella sua integrità.
Si è poi, andati avanti analizzando i 30 minuti con i differenti metodi di calcolo delle distanze.

L'immagine seguente rappresenta il clustering gerarchico effettuato con la matrice della distanza calcolata con il metodo euclideo di default.

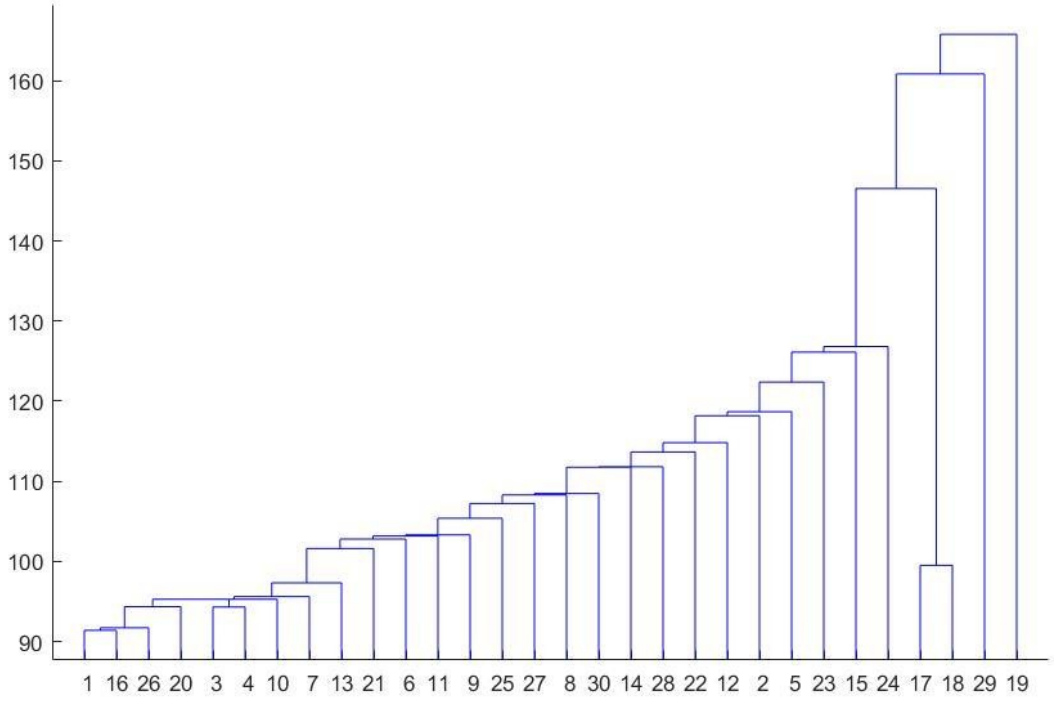

Figura 22

Con l'analisi delle distanze si sono ottenuti miglioramenti con l'utilizzo dei metodi euclideo standard e con quello di Mahalanobis, si riportano di seguito i risultati:

Euclideo standard (seuclidean): c=0.5306

Mahalanobis: c=0.5533

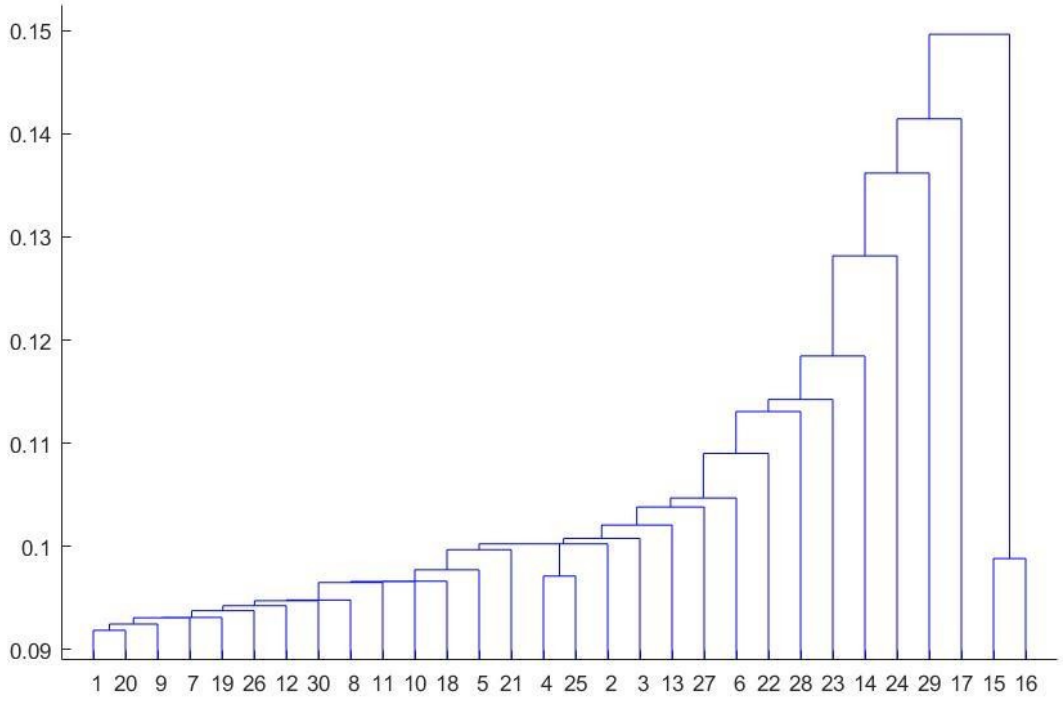

Figura 23

In questo caso si rileva sicuramente un incremento del valore del coefficiente che diventa maggiore di 0.5, ma si è ottenuta una performance migliore sull'altra metà dei dati.

4.2.2 ORE 09

Si prende ora in esame i due gruppi di dati relativi alle ore nove, ovvero il primo insieme che comprende le informazioni inerenti all'arco temporale che va dalle 09.00 alle 09.30 e il secondo insieme che comprende i dati riguardanti la fascia oraria che va dalle 09.31 alle 09.59.

Analizziamo in primo luogo la prima metà dell'ora, il coefficiente di correlazione cophenetica relativo a questi 30 minuti calcolato considerando una matrice delle distanze computata con metodo euclideo ha valore c=0.5598 .

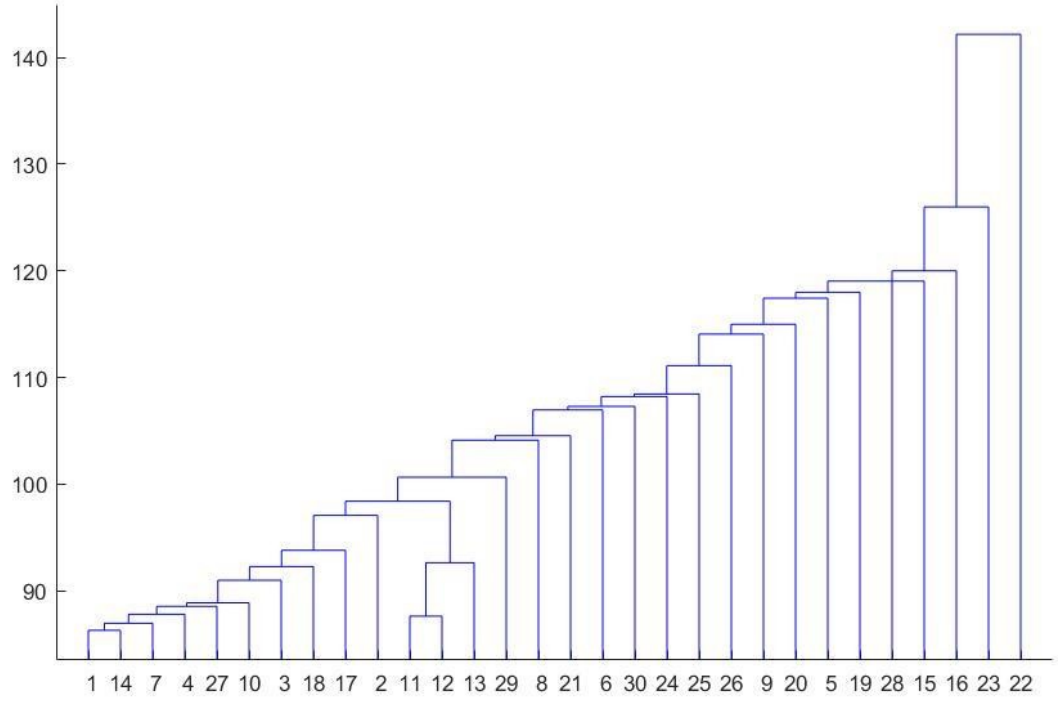

Figura 24

In questo caso abbiamo che il coefficiente ottenuto risulti accresciuto rispetto a ciò che si trovava implementando l'algoritmo sulla totalità dei dati, infatti ora il valore è salito a 0.5598.

Si è andati avanti con l'analisi, implementando l'algoritmo utilizzando i differenti metodi per il calcolo delle distanze.
In ogni caso calcolato si otteneva però un risultato inferiore a quello già ottenuto con il metodo euclideo di default, di conseguenza i dati appartenenti a questa mezz'ora sono rappresentati con il metodo prestabilito.

Analizzando la seconda metà dell'ora, che va dalle ore 09.31 alle 09.59, con il metodo della distanza euclidea si ottiene un coefficiente di correlazione cophenetica, del valore di c=0.6504 .

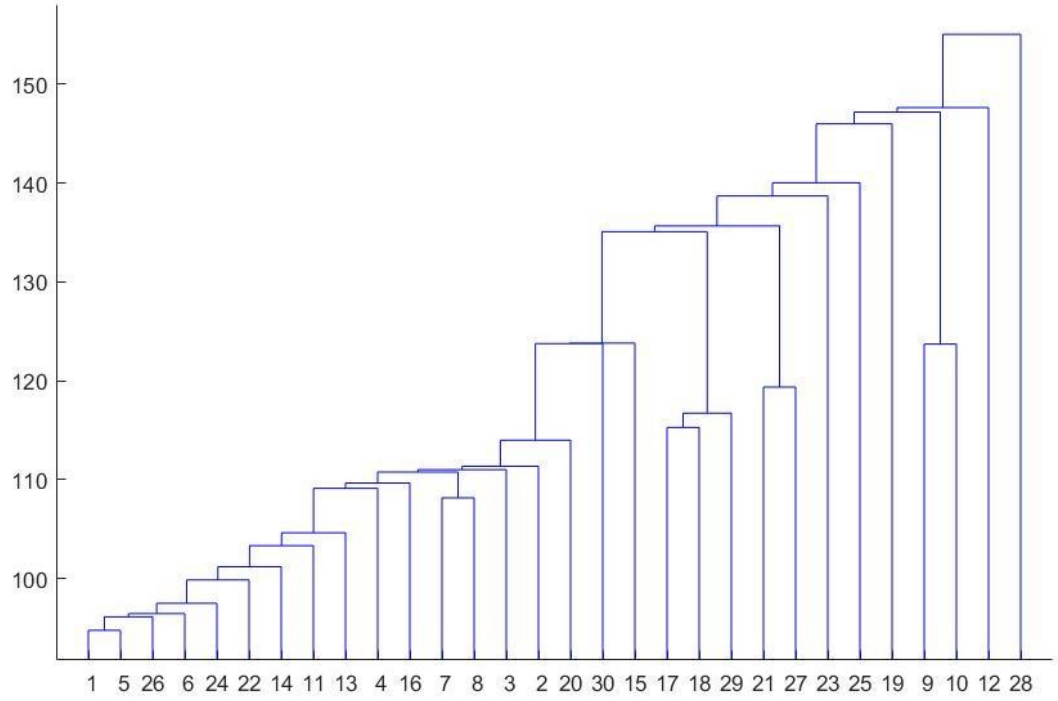

Figura 25

In questo caso il risultato ottenuto è nettamente migliore rispetto al valore del coefficiente avuto con l'applicazione della funzione cophenet () sull'intera ora, infatti si è passati da un coefficiente minore di 0.5 ad uno che è maggiore di 0.6.

Come per le altre fasce orarie abbiamo indagato ulteriormente questo sottoinsieme di dati valutando anche risultati ricavati dall'utilizzo delle diverse tipologie di distanze.

I metodi, che hanno riportato performance migliori rispetto a quella ottenuta con il metodo prestabilito dall'ambiente di calcolo, sono la distanza euclidea standard e il metodo Cityblock, con i seguenti valori:

Euclideo standard (seuclidean): c=0.6540

Cityblock: c=0.6573

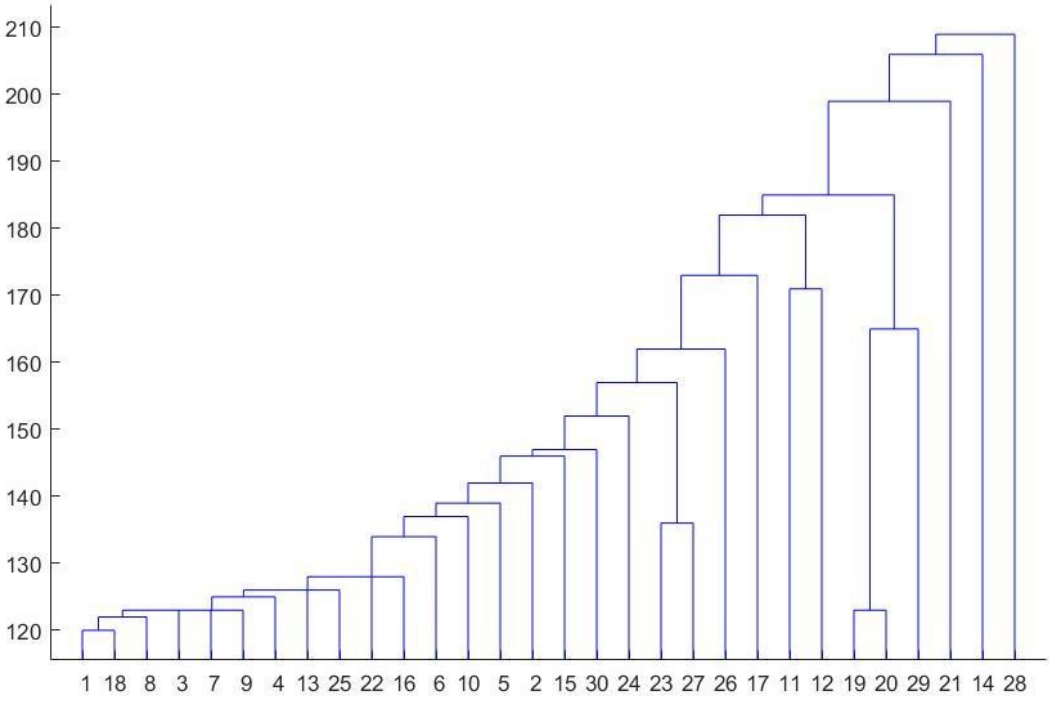

Figura 26

In questo caso otteniamo il miglior valore con la distanza Cityblock, con la quale otteniamo un incremento sia del valore del coefficiente per quanto riguarda l'analisi sulla mezz'ora, ma in entrambe le metà si è ottenuto un risultato migliore dal punto di vista della qualità dei collegamenti creati.

4.2.3 ORE 14

L'ora che andremo ad analizzare adesso è l'ultima che analizzeremo ed è anche l'unica fascia oraria relativa alla giornata del 28/06. Si tratta dell'arco temporale che va dalle 14.00 alle 14.59, sul quale si otteneva un risultato inferiore a 0.5 in tutti i casi analizzati, avendo come coefficiente calcolato con la distanza di default c=0.4744, invece il miglior caso ottenuto sull'ora è stato ottenuto dando in input alla funzione cophenet () una matrice della distanza computata con il metodo di Cosine dà in output un valore di c pari a c=0.4933. Per cui si procederà con l'analisi dei due insiemi da 30 minuti.

Procediamo con l'analisi della prima mezz'ora, per questa prima mole di dati si è ottenuto come risultato del coefficiente di correlazione cophenetico, calcolato con la distanza di default un valore pari a 0.6342.

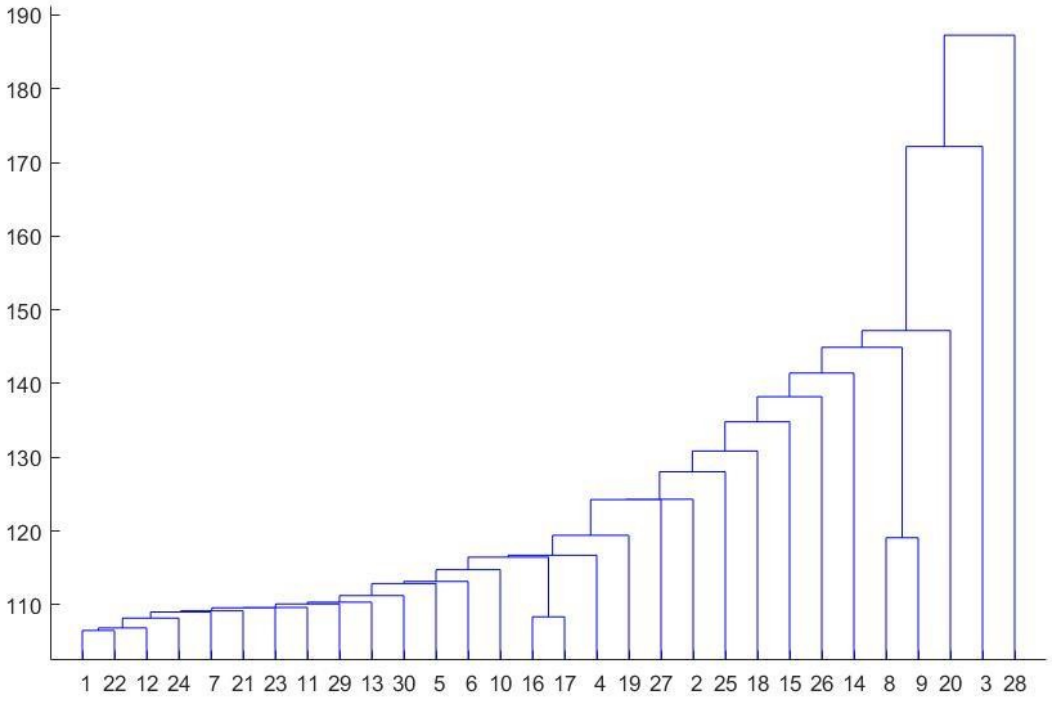

Figura 27

Abbiamo, quindi, un valore superiore a 0.6 già con il metodo di base.
Per cui si è ulteriormente indagata questa partizione per vedere se esistesse un modo per rappresentare in maniera più efficace i dati a disposizione, si è quindi proceduto con l'analisi delle diverse distanze, ottenendo che il coefficiente subisce un incremento se alla funzione cophenet () sono date in input matrici della distanza generate con il metodo euclideo standard e con quello di Chebychev, di seguito si riportano i valori ottenuti:

Euclideo standard(seuclidean) : c=0.6676

Chebychev: c=0.7072

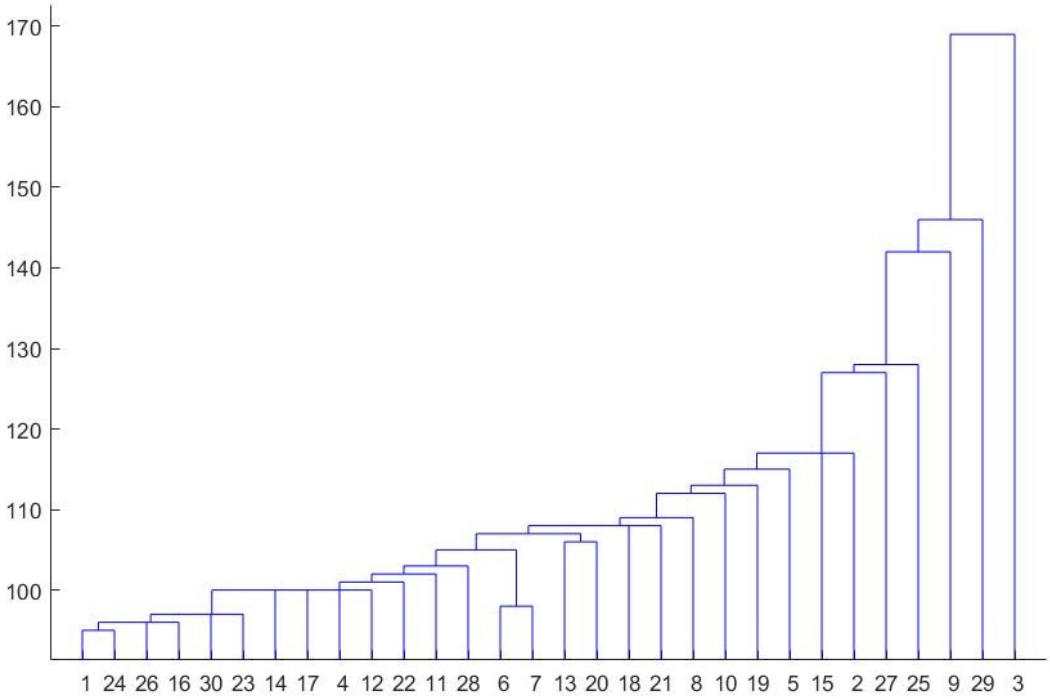

Figura 28

Questo è l'albero gerarchico relativo alla distanza di Chebychev, con la quale si ottiene la miglior rappresentazione dei dati relativi a questa prima mezz'ora.

Procediamo ora con l'analisi dell'ultima mezz'ora, quella che va dalle 14.31 alle 14.59.

Il coefficiente calcolato con la distanza euclidea di default riporta un valore pari a c=0.5416.

In questo caso il coefficiente, risulta comunque superiore al valore che si otteneva dall'utilizzo del clustering sui dati da un'ora, ma rispetto all'altra metà previamente analizzata si nota un calo nella qualità.

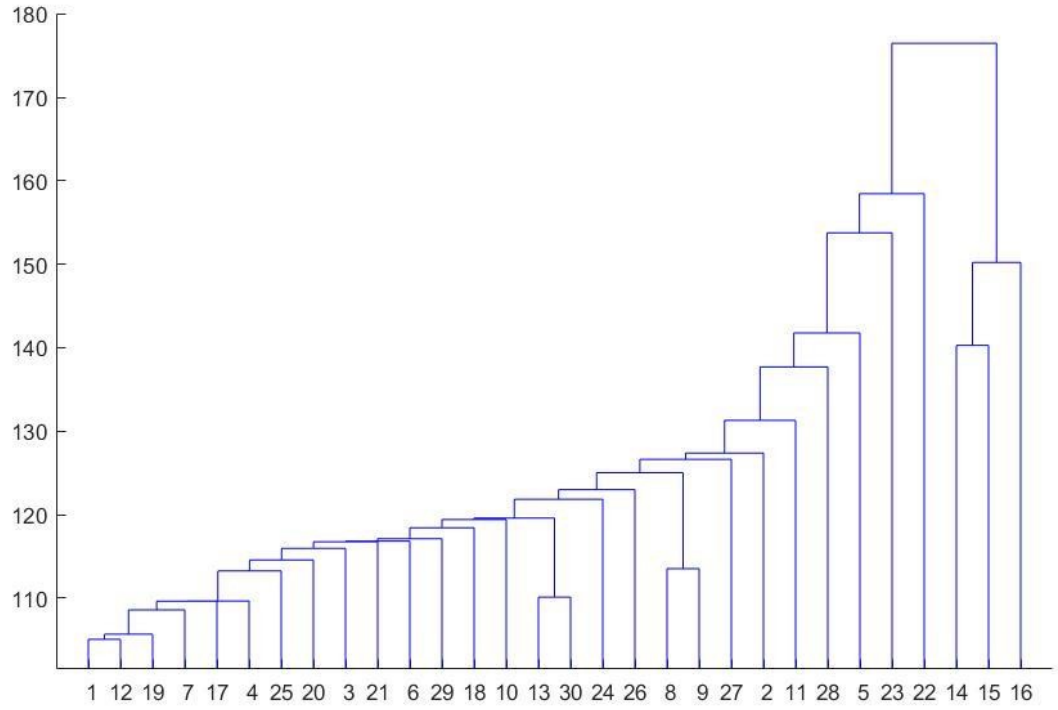

Figura 29

Per cui come fatto per tutti gli altri casi si va ad indagare come questa mole di dati si comporta se viene applicato su di essa lo stesso algoritmo, ma con metodi di calcolo della distanza differenti.

In questo caso solamente il metodo di Mahalanobis faceva sì che il coefficiente di correlazione cophenetica migliorasse, negli altri casi si aveva sempre un valore inferiore a quello ottenuto con il metodo di default.

Infatti, utilizzando Mahalanobis otteniamo un valore pari a:

Mahalanobis: c=0.5514

Questo valore è maggiore di quello trovato senza specificare nessun metodo nella funzione con c=0.5416, ma rimane comunque compreso tra 0.5 e 0.6, che è superiore al 50% ma inferiore al valore trovato facendo questa ricerca sulla mezz'ora precedente a questa.

Di seguito si riporta l'albero gerarchico ottenuto dando in input alla funzione cophenet() la matrice della distanza calcolata con il metodo di Mahalanobis:

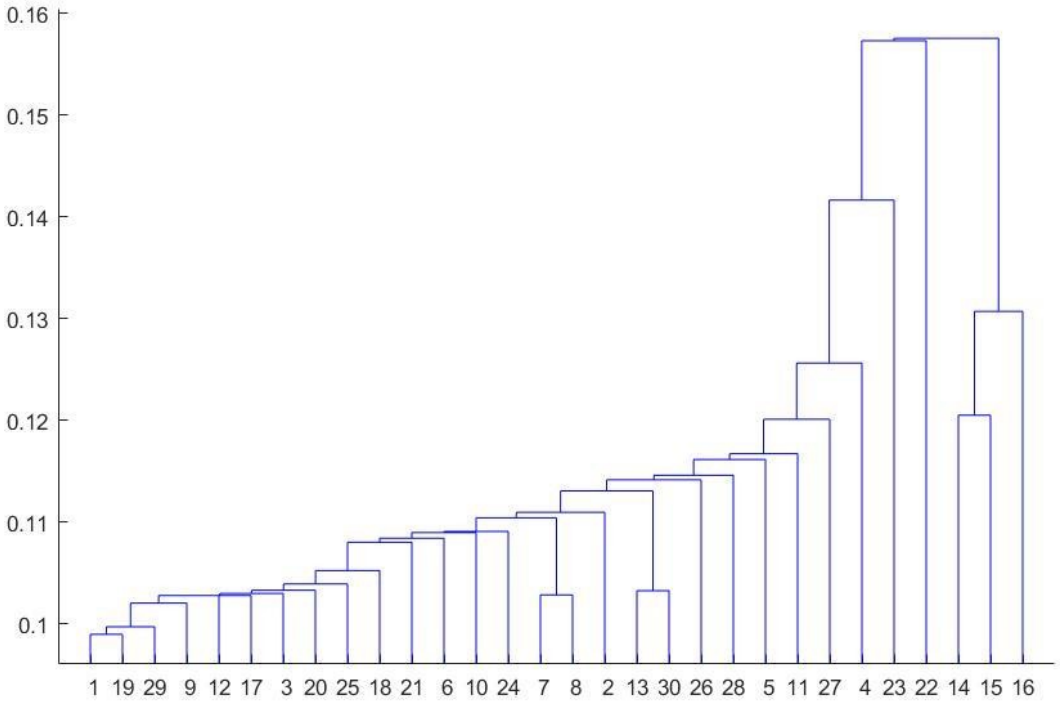

Figura 30

Dall'analisi effettuata su tutti i dati si è notato che in tutti i casi si aveva un miglioramento del valore del coefficiente di correlazione cophenetico rispetto a quello che si otteneva calcolandolo senza specificare alcun metodo nella funzione pdist ().

Però nonostante la maggiore qualità che si andrebbe ad ottenere utilizzando le tipologie di distanze che si sono individuate nel corso di questa ricerca, si è portato a termine il tutto con l'utilizzo della distanza euclidea di default, perché permetteva di mettere a confronto i diversi periodi essendo la stessa per tutti e perché riporta risultati soddisfacenti per la maggior parte delle fasce orarie, in quanto sono solamente 14 su 45 le fasce orarie che riportano un coefficiente inferiore a 0.6; queste rappresentano quindi solo il 33% dei dati analizzati, mentre sono 3 su 45 gli archi temporali che presentano con la distanza euclidea un valore di c inferiore a 0.5, quindi siamo intorno al 6% del totale.

4.3 DISCUSSIONI

Dopo l'analisi di questa partizione di dataset, abbiamo che i dati sono stati accorpati in un numero fisso di cinque cluster, come predefinito dalla ricerca.
L'analisi dei flussi è considerata tenendo conto dei percorsi più frequenti nelle categorie introdotte precedentemente.
Quest'ultima viene effettuata prendendo come riferimento una posizione all'interno di una categoria e successivamente vengono analizzate le scelte di percorso dei clienti, per osservare con quale frequenza una delle possibili direzioni limitrofe è scelta in favore di un'altra.
Dopo una prima osservazione generale delle traiettorie vengono considerati nel dettaglio i percorsi peculiari dei consumatori, in modo da cercare di identificare il comportamento di essi e cercare di estrapolare informazioni sulle categorie di prodotti verso le quali ogni cluster è più frequentemente attirato.

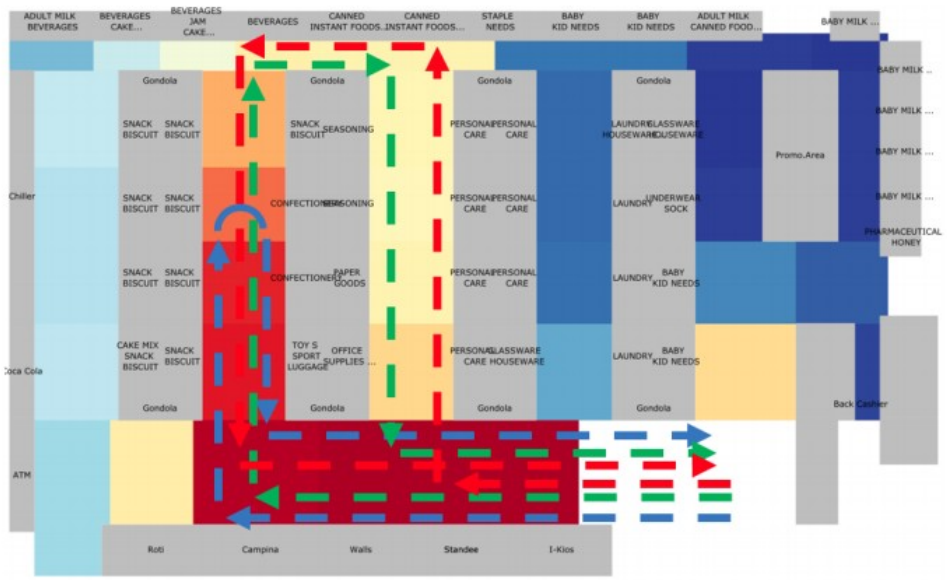

Figura 31

I cluster ricavati dall'analisi dei flussi sono molto utili da un punto di vista del marketing, infatti le traiettorie dei clienti riflettono le preferenze di ogni singolo gruppo individuato con l'algoritmo.
Questo genere di dati è poi messo a disposizione degli imprenditori e dei manager che dai flussi, associati ai tempi di permanenza rilevati e al tipo di prodotti che sono posizionati nella zona considerata riescono a comprendere che tipo di cliente si trova nella propria attività commerciale e a posizionarli all'interno di un cluster specifico.

5. CONCLUSIONI E SVILUPPI FUTURI

L'obiettivo principale di questa ricerca è stato quello di comprendere il comportamento dei consumatori e il loro processo di decisione d'acquisto, in modo da cercare di massimizzare l'impatto che un prodotto può avere sulle persone e quindi aumentare le vendite e far incrementare il profitto dell'attività.
Infatti, l'analisi dei dati provenienti da un ambiente retail e in particolar modo l'analisi delle traiettorie ricavate sono fondamentali per capire e migliorare le strategie di marketing.
La conoscenza dell'andamento tipico d'acquisto degli specifici cluster, può aiutare i manager a comprendere quale sia il miglior layout del negozio, che permetta di ottenere una performance migliore sia in termini economici ma anche in termini di awareness, ovvero che permetta al cliente di avere un'esperienza d'acquisto soddisfacente e di conseguenza associare il negozio con qualcosa di positivo, procurando all'attività anche una promozione passiva mediante l'interlocuzione in un ambiente familiare.
Un'altra applicazione manageriale di questi dati è quella che ci permette di ottimizzare la pubblicità, conoscendo le persone che sono solite frequentare l'ambiente commerciale è possibile individuare luoghi geografici in cui la promozione risulti essere più efficace poiché si tratta di zone frequentate da persone potenzialmente interessate al servizio o bene offerto.
È fondamentale anche per la pubblicità interna al negozio, ovvero qual è il luogo in cui un prodotto riesca ad avere la migliore performance, come ad esempio il posizionamento di prodotti che attirano l'attenzione nei pressi di luoghi in cui c'è una maggiore probabilità di creazione di un assembramento.

5.1 SVILUPPI FUTURI

Ulteriori ricerche su questo argomento possono essere dedicate al miglioramento di questo approccio, mediante l'utilizzo di un insieme di dati più ampio e conducendo l'analisi basandosi su altri indicatori.
Si potrà, magari, sviluppare una tecnologia di ottimizzazione che riesca a supportare il processo decisionale, come il posizionamento degli scaffali o dei prodotti stessi oppure rendere qualche processo automatizzato.
Si potrebbe anche ipotizzare uno studio più approfondito delle traiettorie per comprendere il comportamento di consumatori, però non con il fine economico ma con lo scopo di studiare i percorsi e riorganizzare il layout o creare protocolli che evitino o quantomeno riducano il traffico e gli ingorghi all'interno delle corsie. Questo tipo di ricerca potrebbe avere anche un'applicazione economica poiché il diminuire dell'attesa potrebbe sì rendere vani i prodotti posizionati strategicamente in determinati punti ma allo stesso tempo potrebbe accrescere la soddisfazione dei clienti e la loro opinione del locale, andando a creare una potenziale flessione positiva nell'andamento finanziario dell'attività commerciale.

BIBLIOGRAFIA

[1] K. Ducatel, M. Bogdanowicz, F. Scapolo, J. Leijten, J.-C. Burgelman, Scenarios for ambient intelligence in 2010, Office for official publications of the European Communities Luxembourg, 2001.
[2] M. Sturari, D. Liciotti, R. Pierdicca, E. Frontoni, A. Mancini, M. Contigiani, P. Zingaretti, Robust and affordable retail customer profiling by vision and radio beacon sensor fusion, Pattern Recognition Letters 81 (2016) 30–40.
[3] D. Liciotti, E. Frontoni, A. Mancini, P. Zingaretti, Pervasive system for consumer behaviour analysis in retail environments, in: International Workshop on Face and Facial Expression Recognition from Real World Videos, Springer, 2016, pp. 12–23.
[4] D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, V. Placidi, Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network, in: International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage, Springer, 2014, pp. 146–157.
[5] J.-G. Lee, J. Han, K.-Y. Whang, Trajectory clustering: a partition-and group framework, in: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, 2007, pp. 593–604.
[6] S. Lloyd, Least squares quantization in pcm, IEEE transactions on information theory 28 (1982) 129–137.
[7] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, in: ACM Sigmod Record, volume 25, ACM, 1996, pp. 103–114.
[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, volume 96, 1996, pp. 226–231.
[9] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, in: ACM Sigmod record, volume 28, ACM, 1999, pp. 49–60
[10] W. Wang, J. Yang, R. Muntz, et al., Sting: A statistical information grid approach to spatial data mining, in: VLDB, volume 97, 1997, pp. 186–195.
[11] Adam, Dillon C., et al. "Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong." *Nature Medicine* 26.11 (2020): 1714-1719.
[12] Laato, Samuli, et al. "Unusual purchasing behavior during the early stages of the COVID-19 pandemic: The stimulus-organism-response approach." *Journal of Retailing and Consumer Services* 57 (2020): 102224.
[13] White, Ryen W., and Eric Horvitz. "Cyberchondria: studies of the escalation of medical concerns in web search." *ACM Transactions on Information Systems (TOIS)* 27.4 (2009): 1-37.
[14] Hair, Joseph F., et al. "When to use and how to report the results of PLS-SEM." *European business review* (2019).
[15] https://it.mathworks.com/help/stats/hierarchical-clustering.html#bq_679x-4
[16] M. Paolanti, M. Sturari, A. Mancini, P. Zingaretti, E. Frontoni, Mobile robot for retail surveying and inventory using visual and textual analysis of monocular pictures based on deep learning, in: 2017 European Conference on Mobile Robots (ECMR), IEEE, 2017, pp. 1–6.
[17] D. Buzan, S. Sclaroff, G. Kollios, Extraction and clustering of motion trajectories in video, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 2, IEEE, 2004, pp. 521–524.

ELENCO DELLE IMMAGINI

RINGRAZIAMENTI

Per concludere, non posso esimermi dal ringraziare tutte le persone che sono state coinvolte nella realizzazione di questo progetto, dal Professor Zingaretti, a Marina e Marco che mi hanno aiutato e guidato passo passo nello studio di questo argomento.
Un grazie di cuore va ai miei genitori e a mio fratello Samuele che mi hanno supportato e sopportato in questo percorso, anche a tutta la mia famiglia e alle mie amiche Beatrice e Dalilà che spesso hanno dovuto sentir parlare, a loro malgrado, di derivate, funzioni e molto altro.
Non posso non citare in tutto questo le mie compagne di banco Alessia, Matunikka e Viviana con le quali ho passato molte ore studiando, chiacchierando e cercando di combattere l'ansia preesame nel modo più ironico possibile.
Ed infine volevo ringraziare me, per non essermi arresa ed aver portato a termine questo percorso che si è rivelato a tratti molto difficile, ma che mi ha aiutato a crescere e ha contribuito in parte a farmi diventare la persona che sono oggi.